INSEAD The Business School for the World®

Working Paper
2019/32/TOM
(Revised version of 2018/54/TOM)

Pricing and Job Allocation in Online Labor Platforms

Victor Araman
American University of Beirut, va03@aub.edu.lb

Andre P. Calmon
INSEAD, andre.calmon@insead.edu

Kristin Fridgeirsdottir
London Business School, kristin@london.edu

History: Updated in June 2019.

Problem Definition: We model, analyze, and optimize the operations of an online labor platform that matches jobs to workers. The arrival of jobs and workers to the platform is stochastic and the job processing time is random. The platform chooses the fees per job and assigns jobs to workers with the goal of (i) maximizing platform revenues, (ii) minimizing the unpredictability in workers' profits, and (iii) minimizing any delay in processing the incoming jobs. Workers are sensitive to the revenue they make in the platform and, therefore, the worker arrival rate depends on the platform's pricing and job allocation strategy.
Academic/Practical Relevance: We contribute to the fields of online platform operations and stochastic systems. The asymptotically optimal policy that we introduce is simple and practical. Our analysis provides new insights into the management of uncertainty in online platforms and how uncertainty impacts objectives (i), (ii), and (iii) above.
Methodology: We propose a continuous-time stochastic model that describes the mechanics of the platform. We introduce a policy that sequentially allocates jobs to workers in a rotating manner. The optimal parameters of this policy are \corrected" versions of the fluid solution. Then, through an asymptotic analysis, we prove that this policy is optimal as the platform scales. Finally, we examine the policy's performance through a discrete event simulation.
Results: We introduce a policy class called Uniform Allocation (UA) and provide an analytical characterization of the platform's behavior and performance under this policy class. Then, we design a UA policy that simultaneously optimizes objectives (i), (ii), and (iii) as the system scales. We obtain insights into the UA policy's behavior through a discrete event simulation and find that it leads to similar profits but much lower worker income variability compared to other policies.
Implications: Our proposed pricing and assignment policy align objectives (i), (ii), and (iii) above. From an academic standpoint, we demonstrate how the platform's revenue maximizing pricing and allocation policies can also act as a form of risk pooling. From a practical standpoint, we determine that platform revenue maximization is not incompatible with predictable profits and stable schedules for workers.

Key words : Sharing Economy, Pricing, Scheduling, Queuing, Stochastic Systems.

1. Introduction

Online labor platforms (e.g. Amazon Mechanical Turk, Appen, Uber, Rev.com, etc.) employ millions of workers worldwide (Kässi and Lehdonvirta 2018). Such platforms match skilled workers (the supply), with customers requesting certain tasks to be completed effectively (the demand). For customers, platforms offer convenient on-demand services with minimum fixed costs, transaction costs, and search costs. For workers, platforms provide schedule flexibility and a source of supplemental income with no long-term commitment.

However, double-sided labor platforms face complex operational challenges as they try to sustain an effective matching of supply and demand; they need to operate profitably while ensuring that the needs and constraints of consumers and workers are met. In particular, platforms need to process their customers' job requests in a timely and adequate manner as well as provide workers with an attractive and worthwhile stream of jobs.

Managing the worker (supply) side of the platform is particularly difficult; job availability can be unpredictable or below workers' expectations and income may be inconsistent (Berg et al. 2018). These issues may deter workers from using the platform, thus making job availability even more inconsistent and high-quality service harder to achieve.

Within this context, our research explores how a two-sided labor platform's pricing and job allocation strategies attempt to mitigate such behavior. In particular, we model and analyze the relationship between a platform's pricing and job allocation policy and consider a platform that maximizes an objective comprised of three components, (i) platform's revenues, (ii) workers' welfare measured as the average variability of worker earnings, and (iii) customers' dissatisfaction cost in terms of delays in processing incoming jobs.

Typically, workers are assumed to be driven by the average income generated on a platform. Our approach towards worker welfare is novel because it explicitly considers the *variability* of workers' earnings. As for the platform's control levers, part of the relevant literature, in particular with respect to ride-hailing, have considered dynamic pricing mechanisms to make up for the mismatch between supply and demand. Consistent with the rest of the literature and with the practice of various platforms, we assume that the pricing decision is static, a constant fee (or a fixed percentage) that the platform charges the worker for each transaction. In that context, another lever available for the platform is the job allocation policy, which controls how and when jobs are allocated. This policy would need to balance between the variability of worker revenue and workers' response time to jobs.

Optimizing the platform's objective requires managing multiple sources of risk primarily related to worker availability and workload. To study the relationship between these sources of risk and the

platform's objective we take a queuing-theoretic approach where workers and jobs arrive stochastically; workers *engage* for a certain amount of time on the platform and jobs take a random amount of time to be processed. Workers are strategic in the sense that their decision to join the platform is sensitive to the overall revenue they expect to generate from the platform.

This setting can be modeled as a dynamic stochastic optimization problem. Finding the optimal pricing and dynamic job allocation policy in this setup is intractable, making it difficult to obtain theoretical and managerial insights. Even under standard queuing assumptions (such as Poisson arrival rates and Exponentially-distributed processing times), analyzing such platform performance is akin to examining a network of multi-server queues where the number of servers endogenously changes throughout time. This analysis is even more challenging due to the strategic behavior of the workers.

In order to obtain both tractability and insight, we introduce the so-called *Uniform Allocation (UA)* policies. These policies are static but practical as well as analytically tractable and specifically designed having in mind worker welfare. UA policies assign jobs to workers in a uniform manner by restricting the number of possible *active* workers. The policy replaces natural irregularity (resulting from supply and demand uncertainties) with a short delay in the start time of the worker allocation process. We design a policy in this class that achieves objectives (i), (ii) and (iii) when both supply and demand markets are large (asymptotic analysis). When supply and demand are of moderate size, any improvement in worker welfare is at the expense of maximal matching of supply and demand. UA policies allow for an explicit characterization of the trade-off between worker welfare and the other objectives of the platform.

Before detailing our main contributions, we present examples that motivate our research.

1.1. Practical Motivation: Microwork, Delivery, and Ridesharing Platforms

The first type of labor platform that motivates our research is online "microwork" platforms, such as Samasource, Appen, and Clickworker. These distributed labor platforms are at the core of the emerging "AI Economy" and companies such as Google, Uber, Baidu, Microsoft, and Amazon outsource manual image classification, content filtering, data cleaning, and text translation and transcription tasks to such platforms. These types of tasks are simple, standardized, only take a few minutes to complete, and are time sensitive[1]. In such microwork platforms, workers usually have a job queue and are paid per job. Platforms like Clickworker, CloudFactory, Appen, and Samasource dispatch jobs directly to workers (Vakharia and Lease 2013).

[1] For example, Uber drivers must take a picture of themselves when they log in to the app to verify their identity. When Uber's facial recognition algorithm fails to identify a driver's picture, the identification job is forwarded to FigureEight, who then sends the job to a worker on its platform that will compare the driver's identity with the photo stored by Uber (Gray and Suri 2019).

The other type of platform that motivates our work are delivery and ridesharing platforms such as Postmates, Uber, Lyft and Grab. Recent research reveals that although workers value the flexibility offered by such platforms (Chen et al. 2017), these platforms (at least how they are currently managed) are not necessarily a steady and viable source of income for workers (Daniels and Grinstein-Weiss 2018). Delivery and ridesharing platforms usually allocate jobs directly to workers and, in certain situations, might allow the worker to have a few jobs in its queue. Consistent with recent literature such as Taylor (2018), we do not model the fact that workers in such platforms move between different locations.

1.2. Summary of Main Contributions

The main contributions of the paper are described below.

We introduce an infinite-horizon analytical model of a two-sided labor platform based on continuous time dynamics, which incorporates the platform's triple objective (revenues, customer satisfaction, and worker welfare), the utility function of workers, the uncertainty of job availability and workers, and uncertain job processing time. We model customer satisfaction as the average sojourn time of jobs in the platform and worker welfare as the standard deviation of the number of jobs workers receive during their time on the platform. We assume that workers decide to join the platform based on their expected utility that incorporates the expected income rate on the platform and its predictability. If they join, they spend time T on the platform and, during this time, accept the jobs allocated to them. This framework allows us to explicitly model the interconnection between the platform's pricing policy, job allocation policy, and objective.

We propose a class of simple and practical policies called Uniform Allocation (UA) policies that allocate jobs to workers in a rotating manner and use a "worker buffer" to control the uncertainty in the system. We prove that, for this policy class, we can obtain analytical expressions for platform revenues, for the average job sojourn time, and for the variability in worker revenues. Then, we perform a large-scale system analysis and accordingly design a UA policy that is asymptotically optimal. This analysis shows that a *corrected* fluid solution will be required to offset the impact of uncertainty.

From a methodology point of view, we contribute to the literature of stochastic systems and large-scale operations by demonstrating that, in two-sided platforms, a balanced loading setting is economically optimal and causes limited congestion and more interestingly, little irregularity effects. Furthermore, our model under UA policies allows us to analytically describe the cross-impacts of uncertainty on the platform. Specifically, we derive expressions that describe how, under a UA policy, the wait time for jobs is directly connected to the non-uniformity of workers' schedules. Based on this analysis, we introduce an optimization problem that allows a platform manager to optimize the parameters of UA policies.

Finally, we examine the behavior of the UA policy using discrete event simulation. We compare the UA policy with a random allocation (RA) policy and Shortest Queue (SQ) policy. We find that the UA policy leads to a significantly smaller coefficient of variation (CV) of worker revenues compared to both RA and SQ policies. In addition, the CV of the UA policy is a *decreasing* function of worker arrival rates while the CV of the SQ policy is increasing.

The paper is organized as follows: we review the literature in the next section; in Section 3, we introduce our model and optimization problem; in Section 4, we introduce the UA policy and provide an analytical characterization of the platform's optimization problem under such policy as well as solve the fluid version of the problem; in Section 5, we characterize the asymptotic behavior of a UA policy and obtain an asymptotically optimal solution to the platform's optimization problem based on a corrected fluid solution; in Section 6, we examine natural extensions to our model; Section 7 is devoted to a numerical analysis; in the concluding section, we summarize our findings and list open questions.

2. Literature Review

Our work is related to different streams of literature around pricing and matching of labor platforms. First, our work contributes to the literature on two sided-platform management (see Rochet and Tirole 2006). Recent papers in this streamed examine the matching problem for such platforms in various settings. In the context of ride sharing, Hu and Zhou (2016) and Ozkan and Ward (2017) study the dynamic matching problem. Specifically, Ozkan and Ward (2017) study a continuous-time setting and propose an LP relaxation that yields an asymptotically-optimal policy that performs better than a "closest-driver policy". Hu and Zhou (2016) consider dynamic matching in a discrete-time system and provide conditions under which the optimal policy is a priority rule. Caldentey et al. (2009) and Adan and Weiss (2012) examine the matching of two multi-type sequences on a first-come first-served basis and quantify the matching rate between a type of customer and a type of server.

Recently, Afèche et al. (2017), tackle driver and rider matching in a network matching problem. They consider a fluid model that accounts for the network structure, driver incentives and driver repositioning decisions. The authors then propose a policy that controls the admission to the platform, repositions flows, and also matches drivers to riders locally.

Recent papers have addressed pricing in the context of labor platforms. Cachon et al. (2017) compare different pricing schemes and show that dynamic pricing is better for all stakeholders. Banerjee et al. (2015) study a stochastic queueing model with price sensitive riders and drivers and show like us that static pricing are asymptotically optimal. Bimpikis et al. (2016) examine the platform decisions with respect to the ride pricing and driver compensation across a network. Bimpikis et al. (2016) assume that drivers behave strategically and perform an equilibrium analysis.

Similar to the papers above, we model a two-sided on-demand labor platform with strategic workers where the platform sets the price charged to the workers but, in contrast, we revisit the way the platform should allocate the jobs given its concern with workers welfare.

As opposed to peer-to-peer product sharing (e.g. Benjaafar et al. 2018), our platform shares some common features to so-called freelancing platforms (e.g. Moreno and Terwiesch 2014) but is closer to what is known as on-demand platforms (see, Table 1 in Taylor 2018). In freelancing platforms, jobs are matched with workers that have the proper skill for the job. Conversely, on-demand platforms deal with non-differentiated jobs i.e., any worker can do any job. Taylor (2018) examines an on-demand platform where customers are sensitive to delays and agents that are "independent" can choose to participate in the platform or not. However, while the analysis in Taylor (2018) focuses on the impact of consumer delay sensitivity and worker independence on optimal prices and wages and takes the job allocation mechanism as given, we assume that jobs are homogeneous and examine the design of the job allocation mechanism in order to reduce the unpredictability of the number of jobs allocated to workers during their engagement time. We also assume that the platform charges a fee from workers which is equivalent of charging a commission on the revenue generated by a job. This type of contract has been explored by Hu and Zhou (2017) who show that, under certain assumptions, a commission contract is near-optimal for platforms. Similar to our work, Gurvich et al. (2016) model a setting where a platform optimizes worker pool size and compensation. However, they consider a discrete-time dynamic set-up with less granularity (e.g, no queueing effect), leading to a newsvendor-like relationship between the firm and the workers.

More generally, our work is also related to the literature on pricing and revenue management for a queueing system. This literature often considers a standard single server or multiple server queue and tackle various other issues besides pricing, such as capacity sizing (Savin et al. 2005), lead time quotation (e.g. Çelik and Maglaras 2008, Ata and Olsen 2009), customers with different priorities (e.g. Savin et al. 2005) or the multi-product setting (Maglaras and Meissner 2006). Among these queueing-based pricing literature, the stream of papers that is most relevant to ours is the one using large-capacity systems in the so-called Halfin-Whitt regime (e.g. Halfin and Whitt 1981, Whitt 1992). This type of setting has been used extensively in various applications; in particular in call centers (see Gans et al. (2003) for an overview). In this stream, the closest works to ours are Maglaras and Zeevi (2003) and Maglaras and Zeevi (2005) who provide an equilibrium analysis, determine the demand rate and capacity, and obtain approximations for the optimal solution through "large-capacity asymptotics". The asymptotic analysis they undertake is similar in nature to the one we perform in (where the "heavy-traffic" regime is shown to be optimal from an "economic optimization" point of view). Nevertheless, the system we study (a "wheel"

type matching policy) and its analysis remain different from theirs. Moreover, besides pricing, in a multi-class setting, we look at jobs allocation decisions under fixed total capacity of jobs.

From a practical standpoint, "wheel" policies, similar to the policy we introduce, have been successfully applied in manufacturing[2], King et al. (2016). Scheduling with the goal of generating predictable and smooth production and demand patterns is a key idea in lean manufacturing and is known as *heijunka* in the terminology of the Toyota Production System (Wilson 2015). Thus, our policy can be thought as an application of the concept of *heijunka* to a labor platform.

3. Modeling Framework

We model a two-sided labor platform. The *demand* side of the platform consists of customers that submit jobs to be completed in a timely manner. The *supply* side consists of workers that connect to the platform in search of jobs. The platform *allocates* jobs to workers and charges them a fee per completed job. We now introduce and discuss our model.

3.1. Platform's Decisions

We assume that a worker receives r dollars of revenue for each job she completes on the platform. In return, the platform charges the worker a fee p for each completed job. We focus on the supply (worker) side of the platform and assume the revenue r to be exogenous. The platform manager decides the fee p and the allocation policy π for assigning jobs to workers. We assume that workers are strategic because they are sensitive to the expected revenue they make during their engagement with the platform. As a result, the fee p and the allocation policy π modulate the arrival rate of workers.

In all generality, the platform can choose the fee p and allocation policy π dynamically. The optimization of the resulting stochastic dynamic optimization problem is clearly intractable. Our objective is to examine the problem at a tactical level and identify a simple and well-performing pricing and job allocation policy. This policy should allow us to better understand the implications of the platform's pricing and scheduling decisions on the value generated to the platform, the worker, and the customer. Thus, we restrict ourselves to a set $\mathcal{P}$ of policies that are static and stationary. In other words, the allocation policy itself is not dependent on the state of the system[3]. Moreover, we disregard any transient behavior and perform a steady-state analysis. In Section 7, we compare the policy we propose to other state-dependent allocation policies through a discrete-event simulation.

[2] Product wheel (or rythm wheel) scheduling have been used in chemical and food manufacturing

[3] The policy could well be for example threshold-based but those thresholds could not change with time

3.2. The Demand (Job Arrivals)

We assume that jobs arrive to the platform according to a Poisson process with a rate of μ. The platform allocates immediately a job to a worker. The time it takes a worker to complete a job follows an exponential distribution with rate γ.

Remark 1 *We restrict our analysis to platforms that allocates jobs to workers as opposed to workers selecting available jobs. Many companies do follow such practice (e.g., Uber, Appen, CloudFactory). Moreover, if jobs are homogeneous, such assumption is without loss of generality. It is beyond the scope of this paper to analyze heterogeneous jobs and the implications on the workers' selection process. As for the timing of the allocation. We consider policies that assign a job to workers as soon as the job is received First, this choice leads to a simpler model and more analytical tractability. Second, this reflects the practice of companies such as Appen or Uber, which immediately assign an incoming job to workers, even if all workers are busy finishing previous jobs. However, this immediate allocation is not required in general and we could consider (static) policies that could delay the assignment of a job before allocating it to an available worker.*

3.3. The Supply (Worker Arrivals).

Workers continuously approach the platform for jobs. They are heterogeneous and differ in the amount of time they remain in the platform as well as in the revenue they expect to generate.

Engagement time. Workers that enter the platform will accept any job allocated to them during a period T. We call T the worker's *engagement period*. It is an exogenous random quantity drawn from some distribution $F_T(\cdot)$. We assume that once a job is accepted the worker commits to complete it.

Worker equilibrium arrival rate. We model worker arrivals to the platform as a Poisson process with rate λ. The arrival rate λ is the result of an equilibrium that depends on the worker's utility function and on the platform's choices of fee p and allocation policy π.

The supply dynamics that lead to the equilibrium arrival rate are as follows. A worker decides to join the platform if her expected net utility U during the total time spent at the platform is positive. The expected net utility depends on the expected number of jobs allocated to the worker during her engagement period T, which in turn, depends on the rate at which workers join the platform as well as the fee and allocation policy adopted by the platform. We denote by $\Theta(\lambda, \pi)$ the random number of jobs allocated to a worker in steady state during (a random) engagement period T and we let $N(\lambda, \pi) = \mathbb{E}\Theta(\lambda, \pi)$. We note that the expected value is taken with respect to all the relevant uncertainties, specifically, the arrival of jobs and workers and the engagement time.

We also define the net utility of a worker at the time she is deciding to join the platform as follows:

$$U(\lambda, p, \pi) = (r - p)\, N(\lambda, \pi) - \xi\, \mathbb{E}\hat{T}(\lambda, \pi),$$

where the first term is the expected revenues a worker generates during the engagement period and ξ is the opportunity cost measuring the outside option revenue *rate* available to the worker. The opportunity cost is known to the worker but the platform only knows the distribution of ξ, given by F_ξ. Finally, $\hat{T}$ is the effective time spent at the platform. While T is the time during which the worker accepts jobs, $\hat{T} \geq T$ is the time until the worker processes all the jobs allocated and completely leaves the system. This time is random, and the worker knows its expected value and accounts for it as she decides whether or not to join the platform.

Given an exogenous arrival rate Λ of workers who are contemplating joining the platform, only a fraction of those will decide to join. We denote by $\lambda(p,\pi)$ the steady state arrival rate of workers which is the solution to

$$\lambda = \Lambda\, \mathbb{P}(U(\lambda,p,\pi) > 0) = \Lambda\, F_\xi\left((r-p)\,\frac{N(\lambda,\pi)}{\mathbb{E}\hat{T}(\lambda,\pi)}\right).$$

Under some minimal properties of U and F_ξ, the equilibrium equation above will be guaranteed to have a unique solution. These properties should, in general, hold and are acceptable for the type of policies we are interested in. Note that $N(\Lambda,\pi)/\mathbb{E}\hat{T} \leq N(\Lambda,\pi)/\mathbb{E}T$. Hence, as long as

$$F_\xi\left(r\,\frac{N(\Lambda,\pi)}{\mathbb{E}T}\right) < 1,$$

an equilibrium exists. This inequality should hold under some very general conditions on F_ξ and π. Moreover, $N(\lambda,\pi)$ should be decreasing in λ for a given allocation policy π, where the more workers approach the platform, the less jobs each should expect. The monotonicity of N in λ guarantees that the upper bound $\Lambda\, F_\xi\left((r-p)\,\frac{N(\lambda,\pi)}{\mathbb{E}T}\right)$ has a unique solution. As for the behavior of $\hat{T}$, it depends on the specific allocation; however one would expect that, for reasonable allocation policies, the equilibrium solution is still unique.

For the rest, without loss of generality, we assume that for each static allocation policy $\pi \in \mathcal{P}$, and fee p, the function $\lambda(p,\pi)$ is known and belongs to $(0,\Lambda)$. Equivalently, there also exists an inverse function $p(\lambda,\pi)$ for a given arrival rate λ. We will verify the existence and uniqueness of these functions for the specific policies we are interested in.

Remark 2 *The model we consider assumes that the engagement time T is an exogenous variable. In practice, this time can be more elaborate and endogenous to the system. Specifically, this time can be driven by a number of jobs the worker targets during the engagement time. In that case, the target N would be an endogenous (random) variable and T is the time spent to reach the revenue target. Reality is likely more complex and is a mix of these two behaviors. We discuss the case where the engagement time is driven by the target N in the extensions of section 6 and in the appendix.*

Even though most of the analysis can be done for a random T (except for the main results following Proposition 1), from this point on, we restrict our analysis to the setting in which the engagement time T is constant. We analyze the case of a random T in Section 6. We also explore the implications of a random T on the platform's performance in our numerical analysis. This analysis predicts that our certainty-equivalent-type approach with respect to T, would lead to adequate approximations of the problem, even when T has a high standard deviation.

3.4. Platform's Objective.

The platform has a balanced objective that needs to trade-off between its own revenues, workers welfare and end-customers satisfaction. We describe each component of the platform's objective below. For ease of exposition, we omit T when convenient from the notations.

(i) Platform Revenues. The platform's revenues in steady state are $p(\lambda,\pi)\lambda N(\lambda,\pi)$, where $p(\lambda,\pi)$ is given by the equilibrium equation described in the previous subsection.

(ii) Worker Welfare and Cost of Non-Uniformity. We model the impact of irregularity on workers' welfare as a cost that is proportional to the standard deviation of the number of jobs allocated to a worker in steady state during their engagement time. For an arrival rate λ and allocation policy π, the *cost of non-uniformity* per worker, $\Sigma(\lambda,\pi)$, is given by

$$\Sigma(\lambda,\pi) = b\,\left(\mathrm{Var}[\Theta(\lambda,\pi)]\right)^{1/2},$$

where b is a cost parameter and $\mathrm{Var}[\Theta(\lambda,\pi)] = \mathbb{E}[(\Theta(\lambda,\pi) - N(\lambda,\pi))^2]$.

(iii) Customer Dissatisfaction and Job Sojourn Time. The platform's goal is to offer a high-quality service to end-customers. We measure service quality based on how long a job remains in the platform, from the time it is received until the processing is complete. In practice, depending on the type of platform, jobs can be done sequentially in the order received or in parallel. Workers may even prioritize some jobs over others. In many settings, the job flow and allocation process will impact the efficiency at which workers process jobs, and as a result, a cost of dissatisfaction is incurred. For the sake of anchoring the discussion and obtaining an analytically tractable dissatisfaction cost, we assume the following setting: workers process jobs in a first-come first-served (FCFS) manner and the cost of dissatisfaction is proportional to the expected sojourn time (waiting time plus processing time), in steady-state, of a single-server-queue in which the arrival process results from both the job allocation policy as well as the job and worker arrival rates. We denote by $W(\lambda,\pi)$ the expected sojourn time and the cost of customers' *dissatisfaction* per worker as:

$$C(\lambda,\pi) = c\,N(\lambda,\pi)\,W(\lambda,\pi),$$

where c is a cost parameter. Note that the rate at which the platform processes jobs is $\lambda N(\lambda,\pi)$.

Finally, we revisit the equilibrium equation governing the arrival rate. Given the above model of how jobs are allocated and processed, we conclude that in steady state, it will take each worker, on average, an additional $W(\lambda,\pi)$ beyond the engagement time to process all remaining jobs in her queue. Recall that this time is accounted for by the worker, in the utility function, before the decision to join the platform is made. Hence, the expected effective time is given by $\mathbb{E}\hat{T} = T + W(\lambda,\pi)$ and therefore, the equilibrium arrival rate $\lambda(p,\pi)$ (or equivalently, the price, $p(\lambda,\pi)$) is solution to

$$\lambda = \Lambda\, F_\xi\left((r-p)\,\frac{N(\lambda,\pi)}{T+W(\lambda,\pi)}\right). \tag{1}$$

Remark 3 *We view each worker as a stationary single server queue. Clearly, the waiting time of jobs at a worker's queue will first follow a transient behavior before reaching steady state. But, we do not analyze the transient behavior for tractability reasons and because we believe that the steady state analysis of this queue suffices to demonstrate the general behavior of the above mentioned dissatisfaction cost which could easily apply beyond the specific case of a generic single server queue. We also note that the expected value of the steady state waiting time is an upper bound on the time-average waiting time for each queue. Moreover, as long as the the steady state is eventually reached before the worker leaves the system, we can still write:* $\mathbb{E}\hat{T} = T + W(\lambda,\pi)$. *These two facts will guarantee that our asymptotic results will not be affected by this assumption.*

Putting all the above elements together, we define the profit rate as

$$\Pi(\lambda,\pi) = \lambda\left[p(\lambda,\pi)\,N(\lambda,\pi) - \Sigma(\lambda,\pi) - C(\lambda,\pi)\right],$$

where $p(\lambda,\pi)$ is the solution to equation (1).

The platform's optimization problem is then given by

$$\max_{\pi\in\mathcal{P},\lambda\in[0,\Lambda]} \Pi(\lambda,\pi), \tag{P}$$

Remark 4 *The problem formulation (P) is exactly the same to the one obtained if we would have considered that, instead of the platform incurring the non-uniformity cost, the worker's utility accounted for both the expected revenues and the variability of the jobs allocated by the platform. In this case, the equilibrium equation is:*

$$\lambda = \Lambda\, F_\xi\left(\frac{(r-p)\,N(\lambda,\pi) - \Sigma(\lambda,\pi)}{T+W(\lambda,\pi)}\right). \tag{2}$$

where the constant b measures the workers sensitivity with respect to the predictability of their income. The price is then given by

$$p(\lambda,\pi) = r - \frac{T+W(\lambda,\pi)}{N(\lambda,\pi)}\,F_\xi^{-1}\left(\frac{\lambda}{\Lambda}\right) - \frac{\Sigma(\lambda,\pi)}{N(\lambda,\pi)}.$$

For the rest we use the equilibrium equation (1), and the notion of cost of non-uniformity introduced above. The latter can still be interpreted as an additional element to the workers' utility function.

Remark 5 ***Utilization of the Worker and Stability of the Platform.*** *We define the* workers' utilization *as the average fraction of time workers are busy during their engagement with the platform. The platform manager sets the price and the allocation policy so that the workers' utilization is small enough to ensure a low job waiting time, but not too small to discourage workers from joining the platform. In particular, the cost of dissatisfaction guarantees that any optimal solution ensures that the platform is demand constrained*[4] *, i.e.,* $\mu < \lambda(p,\pi)\,T\,\gamma$.

The optimization problem (P) is, in general, intractable even when we have restricted our feasible policies to the set $\mathcal{P}$. In the next section, we introduce a policy class that is adapted to the problem at stake, allows for analytical tractability and is asymptotically optimal.

4. The Uniform Allocation (UA) Policy

In this section, we introduce and examine a simple and well-performing class of policies, called *Uniform Allocation* (UA) policies. These policies are designed to balance components (i), (ii), and (iii) of the platform's objective and are tractable. Specifically, such policies allow us to measure the impact of uncertainty in the system through what we call the *engagement delay* as well as discover the synergies between the pricing lever and the allocation mechanism. In the next section, we prove that such synergies may lead to an asymptotically optimal performance. We denote the class of UA policies by $\mathcal{P}^{\text{UA}}$.

4.1. Description of a UA Policy

UA policies are *static* and parameterized each by a positive constant κ. This parameter regulates the allocation frequency of jobs to workers in the following way: incoming workers are assigned to slots on a "wheel" with κ fixed slots. Incoming jobs are assigned to the κ slots in a rotating manner (i.e., the wheel "turns"). Hence, a slot (and the worker assigned to it) will receive a job every κ jobs received by the platform. This way, the jobs that are allocated to a worker are spread throughout the worker's engagement period. Figure 1 depicts this policy.

We call a worker *active*, if she is currently assigned to one of the κ slots. If all κ slots are occupied and a worker joins the platform, she will enter a *worker queue* where she waits for a slot to become available and, as a result, the effective start of her job allocation process is delayed. A worker in the queue is said to be in a *passive* state. We denote the time between the worker's arrival to the platform and the time she receives her first job by τ and we call it the *engagement delay*. Passive

[4] If the market is supply constrained then the platform will necessarily have to reject on arrival a portion of the demand that is large enough so that the remaining demand rate is again upper bounded by the supply capacity.

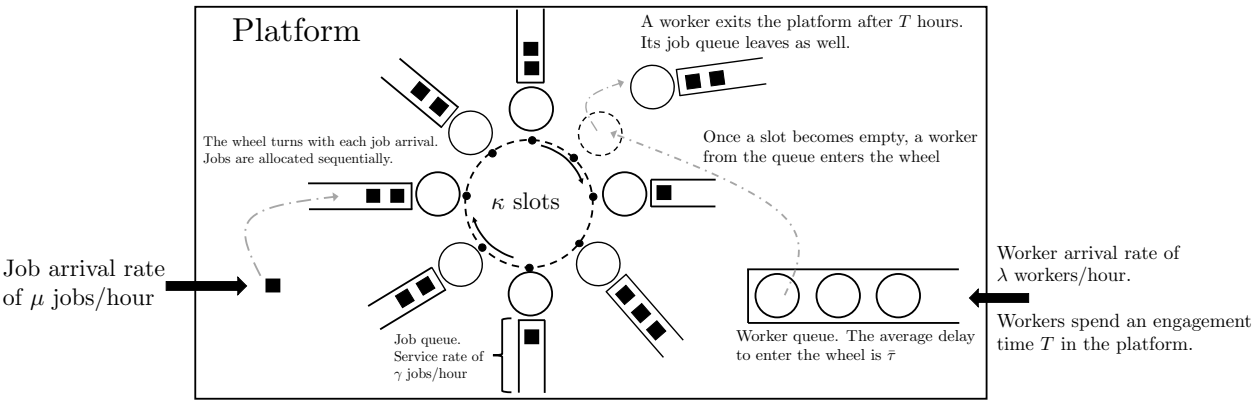


Figure 1 illustration of a UA policy. Workers are represented by circles while jobs are represented by squares.

workers become active in a FCFS manner as slots become available; a slot becomes available as soon as the engagement period of an active worker ends. An active worker is allocated a job every full rotation of the wheel and, therefore, the interarrival time of jobs to an active worker is the sum of κ exponential random variables.

In summary, a UA policy restricts the number of active workers, and hence a worker entering the platform could first experience a potential delay τ during which she receives no jobs. Once the worker is allocated to one of the κ slots, she becomes active for the remainder of the engagement period, $T-\tau$. We call the regularity parameter κ the *wheel size*. An active worker receives jobs at a rate μ/κ. The arrival rate λ is still given by the equilibrium equation (1) and becomes a function of the price p and the wheel size κ that parameterize the allocation policy. We denote the expected number of jobs each worker receives during the engagement time T by $N(\lambda,\kappa)$. As a result, the behavior of the entire platform has been reduced to two parameters: λ (alternatively, p) and κ.

By introducing the regularity parameter, κ, UA policies allow us, to some extent, decouple the arrival of workers and the allocation process. This will be crucial in controlling the variability of the jobs allocation process and in reducing the variability of the jobs sojourn time.

Remark 6 Jobs allocated to empty slots. *When the number of workers on the platform is strictly smaller than κ, the UA policy as defined above might allocate an incoming job to an empty slot. There are different options to deal with this case:*

- *In order to preserve the uniformity of the allocation process, the platform drops the job. The job is then lost and the platform potentially incurs a penalty*[5]*. This approach will be the norm in a supply-constrained market where jobs need to be rejected for the lack of available workers.*

[5] The platform could use some external capacity and outsources job processing. This is typical in online advertising where jobs correspond to viewers and workers to advertisers. In such setting, the platform could direct some of its excess viewers to ad networks. By doing so, they typically generate a lower profit margin.

- *The platform reallocates the job to another slot sampled at random. If the sampled slot is empty, the platform can decide to sample a slot again until the job finds a non-empty slot. This way, in a demand-constrained settings, the platform is guaranteed that no job is lost unless it is empty. By doing so, each worker receives a new stream of jobs that perturbs the uniform allocation process but where the overall variability would remain manageable.*

In all cases, the job is lost if the platform is empty where no worker is in the system. We analyze in Section 6 the two possible approaches introduced in Remark 6. We also discuss their implications on the optimal solution which remains of the same form. For the rest, we ignore the jobs that arrive to empty slots. However, for the numerical analysis we adopt a UA policy with a random reallocation.

Remark 7 *Any allocation policy (even dynamic) generates irregularity (e.g., long or short idle times) in the job allocation process throughout the engagement time. Such inevitable irregularity is "replaced" under the UA policy by an initial delay at the start of the engagement period followed by a (more) regular assignment of jobs to workers. Our assumption is that the worker is strategic in her decision to join the platform, in the sense she is not affected by a short-term behavior (initial delay or irregularity); instead, her decision takes into account a holistic view by calculating its expected utility for the entire interaction with the platform - which accounts for both the overall expected revenues and the corresponding time spent at the platform (see, equation (1)).*

4.2. Engagement Delay

The model we are considering for the labor platform has, under a UA policy lens, a unique feature, whereby the engagement delay plus the time during which the worker is active equals the exogenous constant T. This is not a standard feature of a typical queuing system. However we are able to obtain a tractable formulation of the platform's behavior.

Next, we introduce an expression of the engagement delay τ that will play a central role in quantifying the cost of non-uniformity as well as the cost of customer dissatisfaction. We denote by $\bar{\tau}(\lambda,\kappa) = \mathbb{E}[\tau(\lambda,\kappa)]$ and by $\sigma^2_\tau(\lambda,\kappa) = \mathrm{Var}[\tau(\lambda,\kappa)]$. All proofs are in the appendix.

Proposition 1

i.) For an arrival rate of workers λ and a UA policy with κ slots the engagement delay is

$$\tau(\lambda,\kappa) \stackrel{d}{=} \left[T - \sum_{i=1}^{\kappa} v_i \right]^+,$$

where v_i's are the interarrival times of workers to the platform.

ii.) The expected engagement delay satisfies the following:

$$\textit{For all } \ \kappa \geq 1, \ \ \bar{\tau}(\lambda,\kappa) \leq \bar{\tau}(\lambda,1) \qquad \textit{and} \qquad \bar{\tau}(\lambda,\kappa) \to 0, \ \ \textit{as} \ \ \kappa \to \infty.$$

The main property that enables us to derive Proposition $1(i)$ is that the UA reduces the problem to a single server queue where workers are fulfilled in a FCFS manner.

4.3. Fulfillment Constraint

Under the UA policy with κ slots and an arrival rate λ, each worker will be active for a period $T-\tau(\lambda,\kappa)$. During this period, the worker receives $\Theta(\lambda,\kappa) = \lfloor \hat{\Theta}(T-\tau(\lambda,\kappa))/\kappa \rfloor$ jobs, where $\hat{\Theta}(t)$ is the Poisson process with rate μ counting all the jobs received by the platform during an interval of length t. As a result, $N(\lambda,\kappa) = \mathbb{E}\lfloor \hat{\Theta}(T-\tau(\lambda,\kappa))/\kappa \rfloor$ and what we denote as the *fulfillment constraint*, is given by

$$N(\lambda,\kappa) = (T-\bar{\tau}(\lambda,\kappa))\frac{\mu}{\kappa}. \tag{3}$$

4.4. Cost of Non-uniformity

Next, we obtain a closed-form expression for the cost of non-uniformity.

Lemma 1 *Given λ and κ and assuming that workers have a fixed engagement time T, the cost of non-uniformity is given by*

$$\Sigma^{\text{UA}}(\lambda,\kappa) = b\ Var(\Theta(\lambda,\kappa))^{1/2} = b\left(\frac{\mu}{\kappa^2}\left(T-\bar{\tau}(\lambda,\kappa)\right) + \frac{\mu^2}{\kappa^2}\sigma_\tau^2(\lambda,\kappa)\right)^{1/2}. \tag{4}$$

As mentioned above, there are three sources of uncertainty in this system that may affect the number of jobs received by the worker. The above lemma is quite revealing, as it shows how a UA policy allows us to measure the impact of these three sources of uncertainty through the engagement delay τ of workers. Indeed, the result translates the variability of the number of jobs a worker receives to the steady-state worker's engagement delay distribution (through its average and variance). As for the impact of the uncertainty in the processing times, parameterized by γ, it is implied through the value of the equilibrium arrival rate λ obtained following equation (1).

4.5. Customers Dissatisfaction Cost

The customer dissatisfaction cost is proportional to the expected sojourn time of jobs. Consider the single-server queue of an active worker. Under the UA policy, the arrival process of jobs to the worker's queue follows a renewal process. Given that the processing times of jobs are exponentially distributed, the worker's queue is a $G/M/1$ queue in which the interarrival times of the arrival process is a random variable of the form $\sum_{i=1}^{\kappa} u_i$ where the u_i's are exponentially distributed with rate μ. The sojourn time for jobs in this queue is stated in the lemma below.

Lemma 2 *For a $G/M/1$ queue where the job interarrival rate is the sum of κ exponential random variables with rate μ, the expected sojourn time is finite if and only if $\frac{\mu}{\kappa\gamma} < 1$, (i.e. the worker's utilization is less than one). In this case, the expected sojourn time is given by*

$$W^{\mathrm{UA}}(\kappa) = \frac{1}{\gamma\,(1-\nu(\kappa))},$$

where $\nu(\kappa) \in (0,1)$ is the unique solution to $\nu = \left(1+\gamma(1-\nu)/\mu\right)^{-\kappa}$.

Using this result, the customer dissatisfaction cost is given by:

$$C^{\mathrm{UA}}(\lambda,\kappa) = \frac{c\,N(\lambda,\kappa)}{\gamma\,(1-\nu(\kappa))} = c\,\frac{(T-\bar{\tau}(\lambda,\kappa))\,\mu}{\kappa\,\gamma\,(1-\nu(\kappa))}. \tag{5}$$

Since the platform's objective has a component that is proportional to W^{UA}, any finite-cost policy will require that $\mu/\kappa < \gamma$. As a result, the system will naturally be constrained by the amount of incoming jobs while under-utilizing the workers.

4.6. Platform's Optimization Problem Under the UA Policy

With the non-uniformity and customer dissatisfaction costs in hand, we are ready to formulate the optimization problem (P) for UA policies. Before we do that, we note that $\bar{\tau}$ is decreasing in λ and $\mathbb{E}\hat{T}$ is independent of λ and hence we confirm that the equilibrium equation (1) has a unique solution in λ and the inverse price is given by

$$p(\lambda,\kappa) = r - \frac{\kappa}{\mu}\,\frac{T+W^{\mathrm{UA}}(\kappa)}{T-\bar{\tau}(\lambda,\kappa)}\,F_{\xi}^{-1}\left(\frac{\lambda}{\Lambda}\right).$$

Putting all the above together, in particular, equations (4) and (5), the platform optimization problem under UA policies is given by

$$\max_{\kappa>\mu/\gamma,\,\lambda}\left\{\lambda\,\frac{\mu}{\kappa}\left(p(\lambda,\kappa)-\frac{c}{\gamma\,(1-\nu(\kappa))}\right)(T-\bar{\tau}(\lambda,\kappa)) - \lambda\,\frac{b}{\kappa}\left(\mu\,(T-\bar{\tau}(\lambda,\kappa)) + \mu^2\,\sigma_{\tau}^2(\lambda,\kappa)\right)^{1/2}\right\}. \tag{P^{UA}}$$

Another way to write problem (P^{UA}) is to introduce N as an additional decision variable in the objective function. This formulation emphasizes the indirect control that the platform has on the expected number of jobs delivered to the worker during T. With some abuse of notations with respect to $\Pi(\cdot)$ and $p(\cdot)$, we write that

$$\Pi(\lambda,\kappa,N) = \lambda\,N\left(p(\lambda,\kappa,N)-\frac{c}{\gamma\,(1-\nu(\kappa))}\right) - \lambda\,b\left(\frac{N}{\kappa}+\frac{\mu^2}{\kappa^2}\,\sigma_{\tau}^2(\lambda,\kappa)\right)^{1/2},$$

so that the optimization problem becomes

$$\begin{aligned}\max_{\lambda,\kappa>\mu/\gamma,N}\quad & \Pi(\lambda,\kappa,N)\\ & \bar{\tau}(\lambda,\kappa) = T - N\kappa/\mu \geq 0\\ & p(\lambda,\kappa,N) = r - \frac{T+W^{\mathrm{UA}}(\kappa)}{N}\,F_{\xi}^{-1}\left(\frac{\lambda}{\Lambda}\right) \geq 0.\end{aligned} \tag{$\mathrm{P}_N^{\mathrm{UA}}$}$$

From the fulfillment constraint and the expression of the engagement delay, we conclude that:

$$T - \frac{N\,\kappa}{\mu} = \bar{\tau} = \mathbb{E}\left[T - \sum_{i=1}^{\kappa} v_i\right]^+ \geq T - \frac{\kappa}{\lambda}.$$

If we denote by $\rho \triangleq \lambda\, N/\mu$, the previous inequality implies that $\rho \leq 1$. This inequality reflects the fact that the platform cannot process jobs at a rate higher than the incoming job rate μ. We call ρ the *platform's utilization*, i.e., the fraction of incoming jobs that the platform is processing in steady state. A job is not processed either because no worker is available on the platform at the time of arrival or that the job is allocated to an empty slot (see, Remark 6). Hence, $1-\rho$, is the fraction of jobs that is dropped.

The platform utilization is distinct from the average worker utilization. We denote by u the worker's utilization which is given by $u = N/(T\gamma)$. From Remark 5 we observe that $u \leq \rho$ always holds. The platform can process most incoming jobs while the worker's utilization remains lower (in fact, this is often the case in practice). Given that there is always a probability that the platform is empty (i.e. no workers are available) then ρ will be strictly less than one. Although this is not directly stated in the objective function, many platforms also strive to meet all jobs and, hence, work at a utilization close to one. Interestingly, this is the regime that we show will be optimal for the platform. We finally note that, under a UA policy with reallocation (see, Remark 6) all jobs will be processed (except when no workers are available on the platform). In this case, the effective utilization of the platform is almost one.

Although problems (P^{UA}) or ($\mathrm{P}^{\mathrm{UA}}_N$) have more structure than the original one, (P), they are still "very" nonlinear optimization problems that are difficult to solve even numerically. For example, in the expression of the engagement delay $\bar{\tau}$, κ appears in the limit of a sum. Moreover, the objective is also non-linear in λ, which is the result of a fixed-point expression. We can use a simulation-based optimization to find the optimal values of κ and λ, but a numerical approach does not give us much insight into the dynamics of the platform.

4.7. The Fluid Problem

In order to gain insights into the structure of the solution of problem ($\mathrm{P}^{\mathrm{UA}}_N$) and highlight the natural selection of uniform allocation policies, we examine the case in which the uncertainty in jobs and worker arrivals are disregarded, i.e., workers and jobs approach the platform sequentially in a deterministic and predictable manner and each worker spends exactly an engagement time T on the platform. However, we still assume that jobs processing times are random and follow an exponential distribution with rate γ. We call this setting the fluid problem.

For such fluid problem, and given some N that the platform wants the worker to receive on average during an engagement time T, we set the number of slots on the wheel to be $\kappa^0(N) = \frac{\mu T}{N}$.

We also denote by $\lambda^0(N)$ the corresponding equilibrium arrival rate. The number of slots κ^0 ensures that there will be no engagement delay. Indeed, from the platform's utilization inequality, it must be that $\lambda^0 N \leq \mu$, equivalently, $\lambda^0 T \leq \mu T/N = \kappa^0(N)$. Since, by Little's law, $\lambda^0 T$ is the exact number of workers at any time in the platform (whether passive or active), hence, there will be no engagement delay and the cost of non-uniformity is zero.

As for the cost of customer dissatisfaction, we recall that the worker is modeled as a single server queue that receives jobs sequentially in a deterministic manner at a rate of μ/κ^0. Thus, the worker's queue is a $D/M/1$ queueing system and, as long as $\mu/\kappa^0(N) < \gamma$ or, equivalently, $N/T < \gamma$, the system is stable. Following the same steps as Lemma 2, the next lemma gives the expression of the average sojourn time of a job for such queue.

Lemma 3 *As long as $N < \gamma T$, the expected sojourn time in the fluid problem, under a uniform allocation policy with $\kappa^0(N) = \mu T/N$, is given by*

$$W^{\mathrm{UA},0}(N) = \frac{1}{\gamma}\,\Gamma^{-1}\left(\frac{T\gamma}{N}\right), \tag{6}$$

where $\Gamma(x) = -x\log\left(\frac{x-1}{x}\right)$ is a decreasing convex function on $(1,\infty)$.

The platform's objective function can now be written as a function of λ and N as follows,

$$\Pi^0(\lambda, \kappa^0(N), N) = \lambda N\left(p(\lambda, \kappa^0(N), N) - c\,W^{\mathrm{UA},0}(N)\right).$$

with $p(\lambda, \kappa^0(N), N) = r - \frac{T+W^{\mathrm{UA},0}(N)}{N} F_\xi^{-1}\left(\frac{\lambda}{\Lambda}\right) \geq 0$, and $W^{\mathrm{UA},0}(N)$ given by equation (6).

The fluid version of $(\mathrm{P}_N^{\mathrm{UA}})$ is denoted by (F). Its optimal solution is denoted by (λ^0, N^0). We also let $(\bar{\lambda}, \bar{N})$ be the unconstrained maximizers of Π^0. At optimality, it must be that $N^0/T \leq \gamma$ or else the job sojourn time is not finite. We assume that γ satisfies this assumption. We present the solution for problem (F) below.

Proposition 2 *Consider the fluid problem described above where the interarrival of workers (jobs) is exactly equal to $1/\lambda$ $(1/\mu)$. Assume that F_ξ is such that $x\,F_\xi^{-1}(x)$ is convex. Then, the optimal UA policy that solves (F) has parameters*

$$\kappa^0 = \frac{\mu T}{N^0}, \qquad \lambda^0(N^0) = \min(\mu/N^0, \bar{\lambda}),$$

where

$$N^0 = \begin{cases} \bar{N}, & \text{if } \bar{\lambda}\bar{N} \leq \mu, \\ \arg\max \Pi^0(\lambda^0(N), \kappa^0(N), N), & \text{otherwise.} \end{cases}$$

The above requires that $p^0(\lambda^0, N^0) \geq 0$ which we assume is satisfied throughout. For the rest of the analysis we only focus on the more involved and interesting case where $\bar{\lambda}\bar{N} \geq \mu$. We call this case the *demand constrained* setting[6].

Remark 8 *If the market is supply constrained (with a large μ), then it is technically possible to have $\mu > \bar{\lambda}\bar{N}$ and in this case the optimal solution for the platform is to implement the fluid solution even in the stochastic setting. However, under such a supply-constrained market, customers are likely price sensitive and μ would vary with r. Given the above profit formulation (F), the platform would then start charging a higher fee r which would drop the value of $\mu(r)$, without affecting the proportion of customers that are being served. This process chould continue until at least the constraint $\bar{\lambda}\bar{N} \leq \mu(r)$ becomes binding. Another explanation for this binding constraint is that, in a supply constrained market, $1-\rho = 1-\bar{\lambda}\bar{N}/\mu$ is the rate at which jobs are being rejected or dropped. Eventually, even though the demand is there, this proportion of the market will stop seeking service. The previous analysis is inline with our early assumption that the fee r and the price p are decoupled - a tactic used recently by some online platforms, e.g., Uber, see Garg and Nazerzadeh (2019). However, one could also consider a system where the price charged to workers is proportional to the fare charged to the customers and that the customers are price sensitive. The latter is an interesting avenue to consider but is beyond the scope of this paper.*

The above assumption and Proposition 2 imply that it is optimal for the platform (in a demand-constrained fluid setting) to work at full utilization ($\rho = 1$), and satisfy all incoming jobs. As for the utilization of the worker, it also reaches a maximum value of $u^0 = N^0/(\gamma T) < 1$.

Our starting point for the stochastic analysis is a setting where, in the fluid case, the platform would want to work at full utilization. The convexity assumption of $x\,F_\xi^{-1}(x)$ is done for tractability reasons and many distributions do follow this assumption (e.g., uniform and exponential). However, the analysis in the next section should hold beyond that assumption.

Corollary 1 *We denote by $\Pi^{\mathrm{UA}*}, \Pi^*, \Pi^{0*}$ the optimal profits for problems P, P^{UA} and (F). We have that*

$$\Pi^{\mathrm{UA}*} \leq \Pi^* \leq \Pi^{0*}.$$

This result shows that the fluid optimal profit is an upper bound on the optimal profit under *any* static policy and not only UA policies (refer to, Remark 1). Such bound applies also to policies

[6] Microwork platforms, due to their global labor pool, are usually demand constrained. Berg et al. (2018) conducted a survey of microworkers across multiple platforms and found that a low availability of jobs is a primary concern for more than half of surveyed workers. Recently, Gray and Suri (2019) present an in-depth study of the day-to-day routine microworkers, and both low revenue per job and low job availability emerge as major issues for workers.

that could delay a job before assigning it to a worker as long as it is static and stationary. The proof is straightforward and results from the fact that in the fluid case the cost of non-uniformity is zero and the cost of dissatisfaction is minimal for a given N as long as the arrival process to the worker's queue is deterministic. Finally, N is chosen to maximize the profit at full utilization.

The above result also shows that, at least in the fluid setting, UA policies are optimal among all static policies and motivates the choice of such allocation in a general setting.

Remark 9 The fluid solution for moderate markets. *Although the platform's objective in the fluid problem is an upper bound on the optimal objective in the stochastic case, we cannot expect the "fluid solution",* $(\lambda^0, \kappa^0, N^0)$*, to behave very well. Indeed, the fluid solution is constructed in a way that the allocation spreads jobs evenly while balancing incoming jobs and workers on the platform. Hence, this policy does not leave any room for "error" in the presence of uncertainty. Our objective is then to carefully adjust the fluid solution in order to keep both the cost of non-uniformity and customer dissatisfaction low.*

5. Asymptotic Analysis

In this section, we present an asymptotic analysis of the optimization problem $\mathrm{P}_N^{\mathrm{UA}}$ set in a regime where the arrival rate of jobs and workers are scaled up. We leverage this analysis to simultaneously obtain the asymptotically optimal pricing as well as the design of the UA policy. Next, we define the sequence of scaled problems (P_N^n). Problem (P_N^n) is a scaled version of problem $(\mathrm{P}_N^{\mathrm{UA}})$ parameterized by an integer $n \geq 1$, that is obtained by setting the primitives, $\Lambda^n = n\Lambda$ and $\mu^n = n\mu$. An important feature of our scaling is that the engagement time is not scaled: with $T^n = T$ for all n.

Given that μ is scaled, both the costs of non-uniformity and dissatisfaction as well as the engagement delay τ will be scaled accordingly. In particular, the objective function of problem (P_N^n) is

$$\Pi^n(\lambda^n, \kappa^n, N) = \lambda^n\, p(\lambda^n, \kappa^n, N)\, N - \lambda^n\, \Sigma^{\mathrm{UA}^n}(\lambda^n, \kappa^n, N) - \lambda^n\, C^{\mathrm{UA}^n}(\kappa^n, N).$$

The fulfillment constraint is $\bar{\tau}^n(\lambda^n, \kappa) = T - N\kappa/\mu^n$ and the platform's utilization is $\rho^n = \frac{\lambda^n N}{\mu^n} \leq 1$. We also introduce a new metric, the *congestion factor*, ϱ, which quantifies the load on the platform under the UA policy. The congestion factor is defined as $\varrho = \lambda T/\kappa$, and is the ratio between the average number of workers that are in the system at any point (whether active or passive), and the maximum possible number of active workers κ. As opposed to ρ, ϱ depends on both λ and κ. In a typical multi-server queuing system, ϱ would be exactly equal to the utilization. However, from the fulfillment constraint, in this context $\rho \leq \varrho$. Moreover, ϱ can either be smaller or larger than one. In the fluid case, $\varrho = 1$, but in the general case, it measures whether the platforms balance is leaning toward more workers waiting ($\varrho > 1$) or jobs being dropped or reallocated ($\varrho < 1$).

Furthermore, we define the functions $\Psi(x) = \phi(x) - x\bar{\Phi}(x)$ and $\chi(x) = \bar{\Phi}(x) - \Psi(x)\,(x + \Psi(x))$ on $\mathbb{R}$ where ϕ and Φ are, respectively, the standard Normal pdf and cdf. [7]

The next result sets the appropriate regime for our asymptotic analysis. Interestingly, this regime forces the parameters of the UA policy (κ and λ) to be in a particular form: a *corrected fluid solution*. Moreover, the solution is given as a function of the target N of jobs a worker should expect to achieve during T. In practice, the platform might be driven by a target N to offer the worker; thus the rest of the solution is responsible for implementing such offer in a way to maximize the congestion factor.

Proposition 3 *Suppose that the demand and supply processes are as described above and that the platform is following a UA allocation policy. Also, suppose that both demand and supply rates are scaled as suggested above. Then, for any target of jobs N that the platform could offer, if*

$$\sqrt{n}\,(1 - \varrho^n) \to -\frac{\eta}{T}, \text{ as } n \to \infty, \qquad \text{for some real } \eta, \tag{C}$$

it follows that:

i.) $\lambda^n(N) = \lambda^0(N)(n - \frac{\bar{\tau}(N) - \eta}{T}\sqrt{n}) + o(\sqrt{n})$,

ii.) $\kappa^n(N) = \kappa^0(N)(n - \frac{\bar{\tau}(N)}{T}\sqrt{n} + o(\sqrt{n})$,

iii.) $\rho^n(N) = 1 - \frac{\bar{\tau}(N) - \eta}{T\sqrt{n}} + o(1/\sqrt{n})$,

iv.) $\bar{\tau}^n(N) = \frac{\bar{\tau}(N)}{\sqrt{n}} + o(1/\sqrt{n})$,

as $n \to \infty$, where $\bar{\tau}(N) = \sigma(N)\,\Psi(-\eta/\sigma(N))$ and $\sigma(N) = \sqrt{\kappa^0(N)}/\lambda^0(N)$, with $\kappa^0(\cdot)$ and $\lambda^0(\cdot)$ defined in the fluid setting.

The next proposition shows that under the asymptotic regime when condition (C) holds, the cost of customers dissatisfaction is proportional to the engagement delay.

Proposition 4 *Following the same setting than that of Proposition 3, we have the following:*

i.) the customer dissatisfaction cost is given by

$$C^{\mathrm{UA}^n}(N) - C^{0,\mathrm{UA}}(N) = c\,\beta(N)\frac{\bar{\tau}}{\sqrt{n}} + o(1/\sqrt{n})$$

as $n \to \infty$, where $\beta(N) = \Gamma^{-1\prime}(\frac{T\gamma}{N})$.

ii.) the cost of workers non-uniformity is given by

$$\Sigma^{\mathrm{UA}^n}(\lambda^n, \kappa^n) = \frac{b\,N}{T}\,\frac{\sigma_\tau}{\sqrt{n}} + o(1/\sqrt{n}),$$

as $n \to \infty$, where $\sigma_\tau^2 = \sigma^2\,\chi(-\eta/\sigma)$.

[7] We use hereafter the notation $o(\cdot)$ for two real functions f and g where $g(x) = o(f(x))$ for all x in a neighborhood of x_0 if $g(x)/f(x) \to 0$ as $x \to x_0$.

The first result confirms that (at least under such regime) there is a perfect alignment between improving the performance of the customers (i.e., reducing jobs waiting time) and reducing the delay function of the workers. Moreover, the second result shows that income variability (equivalently, in the allocation process) is also proportional to the variability in the workers' delay function.

In order to state the main theorem, we introduce some notations. We denote by e^0 the elasticity coefficient of the demand function around λ^0, with $e^0 = \lambda^{0\prime} p^0/\lambda^0$, where $\lambda^{0\prime}$ is the derivative of λ with respect to price taken at $p^0 = \lambda^{-1}(\lambda^0, N^0)$. We set $c^0 = (\beta(N)\, N/T)\, c$, and assume in the next result that $p^0 < r$ and $|e^0| > 1$, i.e., the fluid demand is elastic[8]. Finally, we denote by $C^{\mathrm{UA},0}$ the constant $cW^{\mathrm{UA},0}(N^0)$ with $W^{\mathrm{UA},0}(N)$ given in equation (6).

Theorem 1 *Suppose that the demand and supply processes for a single type of engagement are as described above and that the platform is following a UA allocation policy. Also, suppose that both demand and supply rates are scaled as suggested above. Denote by* $\Pi^{0,n}$ *the scaled profit obtained in the fluid setting. Then, at optimality* $N = N^0$ *and condition (C) holds so that the optimal profit is given by*

$$\frac{\Pi^n(N^0)}{\Pi^{0,n}(N^0)} = 1 - \frac{p^0\mu}{p^0\mu - \lambda^0 C^{\mathrm{UA},0}} \frac{\xi^*(\eta)}{T\sqrt{n}} + o(1/\sqrt{n}),$$

as $n \to \infty$, *where* $\xi^*(\eta) = (1 + 1/e^0 + c^0/p^0)\,\bar{\tau}(\eta) + b/p^0\,\sigma_\tau(\eta) - \eta(1 + 1/e^0)$, *with* $\bar{\tau}(\eta) = \sigma^0\,\Psi(-\eta/\sigma^0)$, $\sigma_\tau(\eta) = \sigma^0\,\sqrt{\chi(-\eta/\sigma^0)}$ *with* η *being the minimizer of* $\xi^*(\cdot)$ *and* $\sigma^0 = \sqrt{\kappa^0(N^0)}/\lambda^0(N^0)$.

This result reveals that it is optimal for the platform to deliver to the workers a number of jobs equal to the ones it would have delivered in the fluid case. Asymptotically, the uncertainty is not affecting the expected number of jobs delivered and as a result the utilization of the worker is maximized at teh fluid limit u_0. Moreover, Condition (C) must hold at optimality and thus *the corrected fluid solution* given in Proposition 3 is *asymptotically optimal*. Note that both the average income for the worker and the average profit for the platform will be both lower in the stochastic setting compared to the fluid setting. These losses define the cost of uncertainty that the value of η is minimizing. The impact of the jobs processing time uncertainty is present through both $C^{\mathrm{UA},0}$ and $\beta(N)$. Moreover, this corrected fluid solution ensures a decreasing gap in profits (in the order of $1/\sqrt{n}$) with respect to the fluid setting.

The asymptotic solution presented above is made of the fluid component and a correction term characterized by η, which is uniquely defined by the behavior of the congestion factor.

The observation $\varrho^n = \rho^n \left(1 - \frac{\bar{\tau}}{T\sqrt{n}} + o(1/\sqrt{n})\right)^{-1}$, as $n \to \infty$, implies that such regime can only be achieved under heavy traffic ($\rho \approx 1$). Specifically, as the system scales, it optimally moves toward

[8] A similar assumption was also imposed by Maglaras and Zeevi (2003). We refer the reader to that paper for a brief discussion and illustrative examples of price-demand elasticity.

Table 1 Comparison between $\Pi^n/\Pi^{0,n}$ and the ratio from Theorem 1. We assume $\mu = 20$ jobs/hour, $\Lambda = 10$ workers/hour, $\gamma = 3.5$ jobs/hour, $T = 4$ hours, $b = c = 0.2$, $r = \$1$ and ξ is Exponential with mean 0.8.

n	Π^n	$\Pi^{0,n}$	$(1-\Pi^n/\Pi^{0,n})\sqrt{n}$	$\frac{p^0\mu}{p^0\mu-\lambda^0 C^{\mathrm{UA},0}}\frac{\xi^*(\eta)}{T}$	Error (%)
1	12.26	15.7	0.219	0.202	7.96
10	146.4	157	0.213	0.202	5.13
100	1537	1570	0.209	0.202	3.37
1000	15590	15700	0.207	0.202	2.52

a balanced load at a rate of η/T. Similarly, the utilization gets close to one ($\sqrt{n}(1-\rho^n) \to \frac{\bar{\tau}-\eta}{T}$), while the engagement delay approaches zero making the delivery uniformly spread.

Another important aspect of this result is that the "balanced loading"-type behavior ($\rho \approx 1$) (as well as the rate at which it is reached) is obtained as a result of the optimization and despite the non-linear behavior of the costs of non-uniformity and dissatisfaction. This behavior occurs because as the rate of supply μ^n becomes large, the expected number of jobs, N, becomes much smaller than workload of the platform, $\mu^n T$, received in T. In such regime, by regulating the allocation process, the platform is able to not only keep satisfying the fulfillment constraint, but also take advantage of a risk pooling effect (by increasing the maximum number of active slots available on the wheel (κ^n), simultaneously enabiling it to attract an increasing number of workers (λ^n), without affecting their performance ($\bar{\tau}^n, \sigma_\tau^n$).

Such behavior has been highlighted previously in a queuing context starting with Whitt (1992) and made explicit through economic considerations in Maglaras and Zeevi (2003) and Maglaras and Zeevi (2005). The analysis in these papers is based on a multi-server queuing system in heavy traffic obtained through the Halfin-Whitt regime (see Halfin and Whitt 1981), i.e., by holding constant the probability of *delay*.

Simpy put, Theorem 1 also implies that using a concept of rotating slots is quite effective in reducing the implied uncertainty and generating maximum revenues. In order to illustrate Theorem 1, we consider a numerical example where we scale the system making $\mu^n = n\mu$ and $\Lambda^n = n\Lambda$. We then solve problems (P^{UA}) and (F) for each value of n and compare the ratio $\Pi^n/\Pi^{0,n}$ with the ratio we obtain from Theorem 1 ignoring the $o(1/\sqrt{n})$ term. To make the comparison scale independent, we compare $(1-\Pi^n/\Pi^{0,n})\sqrt{n}$ with $\frac{p^0\mu}{p^0\mu-\lambda^0 C^{\mathrm{UA},0}}\frac{\xi^*(\eta)}{T}$. The results are displayed in Table 1. For moderate values of n the ratio from Theorem 1 is close to the ratio obtained numerically by explicitly solving (P^{UA}) and (F) and, as the system scales, these ratios converge.

To summarize, for systems where the capacity μ is large relative to N, say $\mu = n$, computing the fluid solution together with setting the optimal congestion factor ϱ through the computation of η allows one to solve for the optimal price and the allocation frequency. It also allows one to measure

the system's performance through the utilization ($\rho \approx 1 - \frac{\bar{\tau}-\eta}{T\sqrt{\mu}}$), the irregularity ($\bar{\tau} \approx \frac{\bar{\tau}}{\sqrt{\mu}}$), and the profit ($\Pi \approx \Pi^0(1 - \xi^*/\sqrt{\mu})$). This asymptotic analysis leads to approximations that numerically (see Table 5 are valid for reasonable values of n, making these results valuable in practice.

Together, the results in this subsection demonstrate that, despite its simple structure, a UA policy can achieve the three goals of the platform manager described in the introduction. For adequate policy parameters, as the platform scales the UA policy converges to the fluid profit while simultaneously minimizing job wait times and minimizing variability in worker profits.

6. Extensions

In this section, we propose natural extensions to our model. We do not handle these in full details. For some, we defer the analysis to the appendix. For the rest, we discuss the main approach on how to tackle them.

6.1. Jobs Allocated to Empty Slots

We tackle in this section the different approaches to handle the situation where jobs are allocated to empty slots (see, Remark 6).

6.1.1. Dropout Cost. The first approach is the one where the platform incurs a penalty each time a job is dropped. We denote the unit dropout cost by s. The total dropout cost is an additional cost that the platform incurs as a result of implementing a UA policy. The average total cost rate is given by $\Gamma^{\text{UA}} = s\,(\mu - \lambda N) = s\,\mu\,(1-\rho)$. Indeed, for a given unit of time, the total number of jobs received is μ while $\lambda\,N$ is the number of jobs allocated to workers during that time. The difference are the jobs dropped. From Proposition 3 we conclude that

$$\Gamma^{\text{UA}^n} = s\,\mu\,(\frac{\bar{\tau}-\eta}{T})\,\sqrt{n} + o(\sqrt{n}),$$

as $n \to \infty$. The previous term will be added to the profit and the corresponding asymptotic analysis will lead to a similar form of solution where the penalty cost s will be integrated to the value of η. The steps of the proof would remain exactly the same than for Theorem 1. The result takes advantage of the fact that the penalty cost is zero in the fluid setting.

6.1.2. Random Reallocation Policy. The other more interesting approach discussed in Remark 6, is the one where the platform implements a uniform allocation policy with random reallocation. Specifically, each time the UA policy allocates a job to an empty slot, the platform can reallocate it randomly (and uniformly) to another slot. It iterates this process until the job finds an active worker. By doing so, the platform ensures that the jobs is not lost (unless no workers are available on the platform when the job is received). By applying this to all jobs the platform is guaranteed to work at full utilization.

As a result of such adjusted policy, the workers will receive two streams of jobs the one driven by the uniform allocation and the one resulting from the reallocation policy. The latter stream of jobs can be shown to be a Poisson process. We denote its rate by μ_0. Indeed, every job arriving to the platform has $1-\rho$ probability of finding an empty slot. The reallocation process is a deleted Poisson with a geometric rate equal to $\mu_0 = \mu\,(1-\rho)/\kappa \sum_{i=1}^{\infty}(1-\rho)^{i-1}(1-1/\kappa)^{i-1}$. Simple calculations show that asymptotically,

$$\mu_0^n \sim \mu^n\,(1-\rho^n)/\kappa^n \frac{1}{\rho^n},$$

as $n \to \infty$. We will not attempt to obtain the adjusted formulation of the corresponding asymptotic analysis. Instead, we argue below why the results obtained with a reallocation policy leads to the same form of corrected fluid solution. First, observe that the revenue rate is similar to the one before, except that N is now the aggregate number of jobs allocated during $T-\tau$, from both streams, at a rate $(\mu^n+\mu_0^n)/\kappa^n$. This new rate has marginally increased. Adding this new stream of jobs will also impact the variability with respect to the number of jobs received by a worker. Given that the new stream is a Poisson process independent of the initial UA allocation process, the variance of the jobs allocated increases by a term equal to $(1-\rho)\mu/\kappa(T-\bar{\tau})$. Finally, the expected sojourn time of a job will also increase due to this reallocation stream. Measuring the exact impact of the additional stream is complex because the allocation process is non-Poisson. However, this increase in the expected sojourn time should be smaller than if the (primary) allocation process is a Poisson process with the same rate (i.e., the worker's queue was an $M/M/1$. We do not prove this claim that can be confirmed numerically. It is not hard to show that if we replace the $G/M/1$ queue by an $M/M/1$, then the impact of the new stream on the expected sojourn time is in the order of $1-\rho$ and would represent an upper bound on the real system. The resulting profit rate will be of the same form than Π and the asymptotic analysis would then follow the same way.

6.2. Random engagement time, T

In this section, we consider the case in which the worker has a random exogenous engagement time T with distribution F_T. We define $\bar{T} = \mathbb{E}T$. We use the same notation of the constant T case. We add a subscript T when the quantity is defined conditional on the value of T. For instance, we define $N_T(\lambda,\pi) = \mathbb{E}[\Theta(\lambda,\pi)|T]$ and obtain that $N(\lambda,\pi) = \mathbb{E}_T N_T(\lambda,\pi)$. We still denote the engagement delay by τ. For random T, the expression of τ obtained in Proposition 1 is no longer valid. The value of τ is independent of a realization of T, depending only on the distribution of T.

6.2.1. The Different Ingredients of the Profit Rate. Recall that the equilibrium equation (1) was obtained for a random T. Hence, the revenue rate generated for a given λ and π maintains the same formulation: $\lambda\, p(\lambda,\pi) N(\lambda,\pi)$. Also, note that the worker remains active during

$T-\tau$, the fulfillment constraint holds for each worker, and for a given T: $N_T(\lambda,\pi)=(T-\bar{\tau})\frac{\mu}{\kappa}$. By taking the expected value with respect to T, we recover the same fulfillment constraint as before.

It is not difficult to show that the expected sojourn time remains the same under a random T and that is due to the stationarity assumption[9]. The formulation of the cost in Lemma 2 remains valid. As for the cost of non-uniformity incurred by the platform it is measured as the expected value of the conditional standard deviation $\mathrm{Var}(\Theta_T(\lambda,\kappa))^{1/2}$. Hence,

$$\Sigma^{\mathrm{UA}}(\lambda,\kappa)=b\,\mathbb{E}_T\left(\frac{\mu}{\kappa^2}(T-\bar{\tau}(\lambda,\kappa))+\frac{\mu^2}{\kappa^2}\sigma_\tau^2(\lambda,\kappa)\right)^{1/2}.$$

6.2.2. Approximating the Profit. The difficulty in the setting where T is random is the lack of an expression for τ. In order to circumvent this, we suggest a bound $\hat{\tau}$ on τ obtained following a more stringent model. For that, assume that workers are pre-assigned to specific slots independently of the state of the system. Such approximated system will be less efficient and does not take advantage of risk pooling effect from a centralized queue. However for this system, we can prove that the delay function of the workers is given by a formulation similar to the deterministic case with $\hat{\tau}=[T-V]^+$, where T is random and V is the sum of κ exponentials. The new engagement delay should be stochastically larger than τ. Then, the following inequality must hold

$$\hat{\Pi}(\lambda,\kappa)\le\Pi^T(\lambda,\kappa)\le\Pi^0(\lambda,\kappa),$$

where Π^0 ($\hat{\Pi}$, Π^T) is the profit in the fluid setting (corresponding to $\hat{\tau}$, τ). For an asymptotic analysis, we have to scale T by defining a sequence T^n, that converges to a positive constant $\bar{T}$ as $n\to\infty$. As a result, one can obtain a corrected fluid solution that is asymptotically optimal with $\hat{\Pi}^n/\Pi^0\to 1$ as $n\to\infty$. This implies that this same solution will be asymptotically optimal for Π^T.

6.3. Engagements driven by a target number of jobs

A natural variant of our model is one where the worker engages with the platform until she reaches a target N that is exogenously set. As a result the engagement time in this case is endogenous corresponding to the time it takes to reach N. We call this setting an N-engagement. The worker only joins in this case if the expected utility is positive. Similarly to the case with an exogenous T, under an N-engagement setting, the utility takes into account the engagement time, the number of jobs targeted and an outside opportunity cost. We write that $\lambda=\Lambda\,\mathbb{P}\left((r-p)\,N-\xi\,\mathbb{E}\hat{T}>0\right)$, with $\hat{T}=\tau+\sum_{i=1}^{N}\sum_{j=1}^{\kappa}u_{i,j}+W^{\mathrm{UA}}(\kappa)$, where $u_{i,j}$'s are i.i.d. and represent the interarrival times between jobs. The first term is the engagement delay, the second term measures the (active) time

[9] One might force the value of T to be larger than some $T_0>0$ so that the stationarity assumption remains sensible.

it takes to allocate the N jobs, and finally, the third term is the time (beyond the active time) it takes to complete all remaining allocated jobs. Hence,

$$p(\lambda,\kappa) = r - F_{\xi}^{-1}\left(\frac{\lambda}{\Lambda}\right)\left(\frac{\kappa}{\mu} + \frac{\tau + W^{\mathrm{UA}}(\kappa)}{N}\right).$$

In this more elaborate setting, we obtain again an expression of τ and an asymptotic solution similar to that of Proposition 3 and Theorem 1. These results and their proofs are in Appendix B.

7. Discrete Event Simulation and Numerical Experiments

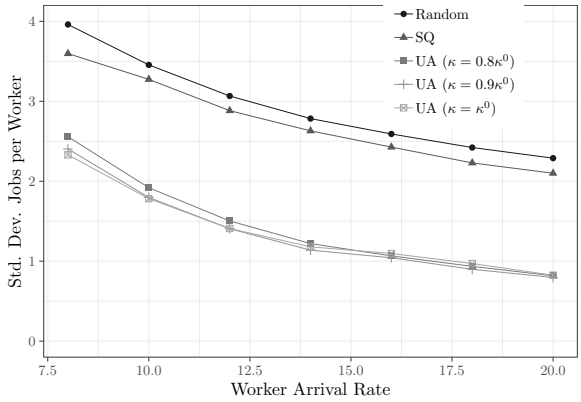
In this section, we examine the performance of the UA policy through a discrete event simulation. Our discrete event simulator was built in the Julia programming language (Bezanson et al. 2017) and simulates each worker and job arrival to the platform, job queues and processing times, and also worker departures. All our code is available at *[link redacted for peer review].* For the UA policy, we assume that if a job arrives to a slot without a worker it is randomly reallocated to other slots until a worker is found, as discussed in Section 6.1.2. In all policies considered, we assume that if no workers are on the platform, the job is lost.

7.1. Comparison with the Random Allocation and Shortest Queue policies

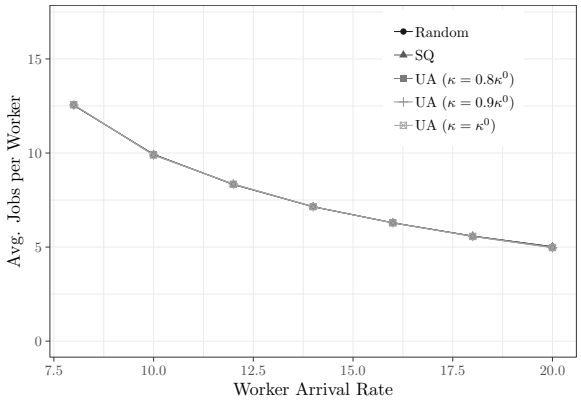
We benchmark the UA policy against the *Random Allocation* and the *Shortest Queue* (SQ) policies for different worker arrival rates. In the Random Allocation policy, incoming jobs are randomly assigned to a worker in the platform. In the SQ policy, incoming workers immediately enter the pool of active workers. Conversely, incoming jobs are immediately allocated to the worker with the shortest job queue (hence, the policy dynamically responds to worker loads). The platform manager's decision is only the fee p. A complete analytical characterization of the platform's objective under the SQ policy is challenging because the arrival rate of jobs to a worker's queue is a function of the queue-length of all workers on the platform.

The results are depicted in Figure 2. The standard deviation of the number of jobs per worker, which we take as a measure of worker welfare, is much smaller in the UA policy than in both the Random Allocation and SQ policies, as shown in Figure 2a, which highlights the main practical feature of the UA policy: it is designed to minimize the variability in the number of jobs allocated to each worker. All three policies allocate the same average number of jobs per worker as displayed in Figure 2b.In our simulations, average worker utilization ranged from about 0.9 to about 0.3.

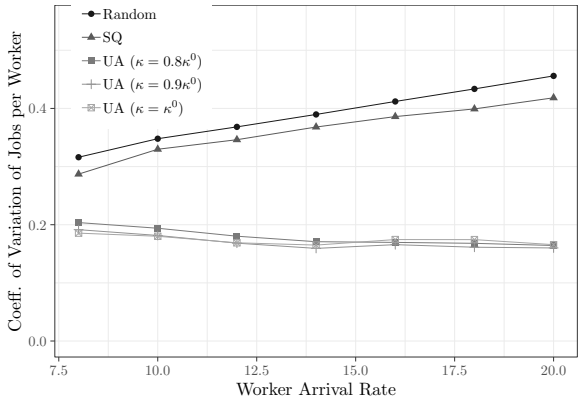
Furthermore, Figure 2c indicates that, while the Coefficient of Variation (CV) of the Random Assignment and of the SQ policy are increasing functions of the worker arrival rate, the CV of the UA policy is a decreasing function of the worker arrival rate. Together, Figures 2c and 2d depict another key advantage of the UA policy: the CV of the number of jobs per worker and the job wait times are *aligned* in the sense that they are both decreasing functions of the worker arrival rate.



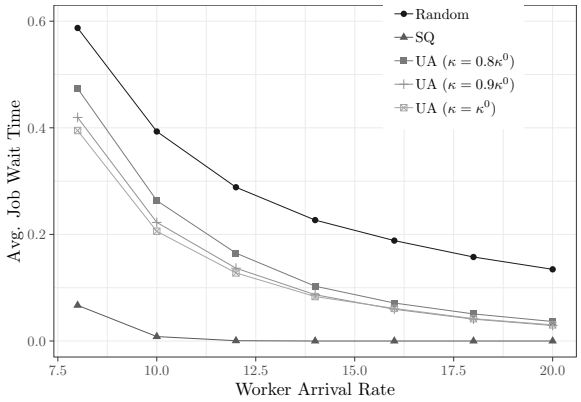
(a) Standard Dev. of jobs per worker



(b) Avg. number of jobs per worker (all curves overlap).



(c) Coeff. of Variation of jobs per worker



(d) Avg. job wait time. (hours)

Figure 2 **Comparison between the UA policy, the Random Allocation policy, and the SQ policy for various levels of λ. We assume $\mu = 100$ jobs/hour, $\gamma = 3.5$ jobs/hour, and $T = 4$ hours. We vary λ between 8 and 20 workers/hour and for each value of λ we simulate the system for 100,000 arrivals of jobs and workers. For the UA policy, we scale the number of slots on the wheel proportionally to the fluid wheel size $\kappa^0 = \lambda T$. We simulate three scalings: $0.8\kappa^0$, $0.9\kappa^0$, and κ^0.**

Finally, as expected, in all policies the job wait times decreases as the worker arrival rate increases, as show in Figure 2d. Also, the wait time UA policy is lower and decreases faster with λ than in the random allocation policy. Since the SQ policy minimizes job wait times by allocating incoming jobs to the SQ, the job waiting time in the SQ policy is significantly lower than in the UA policy.

7.2. Validity of the optimization problem (P^{UA})

We turn our attention to the optimization problem (P^{UA}). Two main advantages of the UA policy are the analytical expressions for job wait times and for the variability of worker revenues, allowing (P^{UA}) to be solved analytically in a tractable manner. However, (P^{UA}) has a few assumptions, such as the steady state of workers queue. In order to evaluate the validity of our theoretical model, we

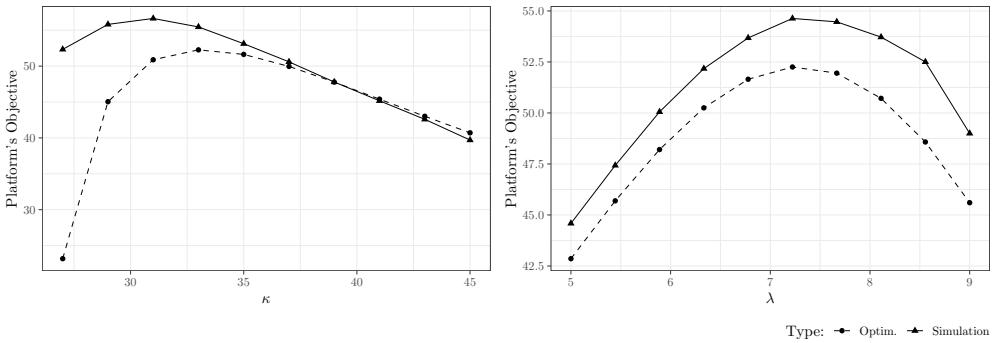


Figure 3 **Platform's theoretical and simulated objective for different values of λ and κ. We assume $\mu = 100$ jobs/hour, $\Lambda = 15$ workers/hour, $\gamma = 4$ jobs/hour, and $T = 4$ hours, $r = \$1$, ξ Exponential with mean $0.8r$. We simulate the system for $300,000$ arrivals of jobs and workers.**

compare the platform's objective function obtained through discrete event simulation and obtained theoretically around the values of λ and κ that maximize (P^{UA}).

The result of this simulation is displayed in Figure (3). For the parameters used in the simulation, the optimal κ and λ that solve (P^{UA}) are $\kappa = 33$ slots and $\lambda = 7.3$ workers/hour. In the first figure, we set $\lambda = 7.3$ and vary κ while in the second figure we set $\kappa = 33$ and vary λ. Although, as expected, the optimal values of κ and λ that maximize the objective in the discrete event simulation the are not the exact maximizers of (P^{UA}), the optimal simulated system parameters are close to the theoretical ones. In fact, a broader numerical search indicates the values of λ and κ that maximize the objective in the discrete event simulation are both within 7% of the values obtained theoretically. More importantly, if we use the values of κ and λ in obtained through (P^{UA}) in the simulated system, the reduction in the platform's objective will be less than 5%. Hence, the theoretical model, despite its assumptions, produces policy parameters with near-optimal practical performance.

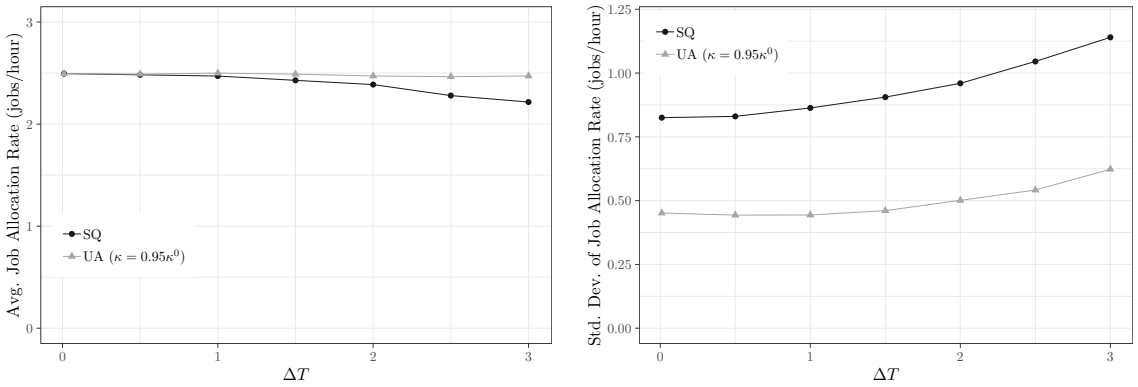
7.3. Sensitivity analysis with respect to γ

As we increase the job processing rate γ, the average job waiting time decreases. However, the effect of increasing γ on the variability of jobs that workers receive is less clear. We investigate this effect through a discrete event simulation and the results are displayed in Table 2. Increasing γ increases the standard deviation of jobs per worker (and the non-uniformity cost) and the average number of jobs per worker. The net effect is a small increase in the coefficient of variability of the jobs workers receive. Conversely, the effect on job wait times is dramatic, leading to an order of magnitude reduction in job wait times.

In addition, as γ increases, the reduction in job wait times outweighs the increase in the variability of jobs workers receive. The net effect is that, at optimality, the platform uses less workers. Hence, p increases and the platform revenues increase, while the average worker revenue decreases.

Table 2 **Sensitivity of simulation for different values of γ. We assume $\mu = 80$ jobs/hour, $\Lambda = 20$ workers/hour, $T = 4$ hours, $r = \$1$, ξ Exponential with mean 0.8, $b = 2$, and $c = 0.2$. We obtain λ, κ, and p by solving (P^{UA}). We simulate 500,000 arrivals of workers and jobs for each γ.**

γ	Avg. jobs per worker	Std. dev. of jobs per worker	CV of jobs per worker	Avg. job wait time	λ	p	Avg. worker rev. per engagement
3.5	8.72	1.69	0.19	0.17	9.14	0.74	2.29
4.31	9.67	1.97	0.2	0.13	8.28	0.8	1.92
5.12	10.5	2.25	0.21	0.097	7.61	0.84	1.68
5.94	10.8	2.36	0.22	0.064	7.39	0.85	1.58
6.75	11.3	2.52	0.22	0.047	7.15	0.87	1.51



(a) Avg. rate of jobs per worker. (b) Std. Dev. of rate of jobs per worker

Figure 4 **Comparison between the UA policy and the SQ policy for random T. We assume that the distribution of T is uniform with support $[4 - \Delta T, 4 + \Delta T]$. We assume $\mu = 100$ jobs/hour, $\gamma = 4$ jobs/hour, and $\lambda = 10$ jobs/hour. We simulate the system for 100,000 total arrivals of jobs and workers.**

7.4. The effect of random engagement time T

In most of our previous analysis we assumed that workers' engagement time T was deterministic. We now explore the effect of random T in the performance of the UA policy. When T is random, incoming workers no longer receive the same amount of jobs. Thus, we measure the cost of non-uniformity as the standard deviation of the *rate* at which workers receive jobs. More specifically, recall from Section 6.3 that $N_T(\lambda, \pi)$ is the random amount of jobs a worker receives for an engagement time T. Then, non-uniformity of the job allocation rate is $\mathbb{E}\mathrm{Var}\left(\frac{\Theta(\lambda,\pi)}{T}|T\right)$, where $\frac{\Theta(\lambda,\pi)}{T}$ is the (random) rate at which a worker receives jobs. For this simulation, we assume that T is a uniform random variable with mean 4 hours and support $[4 - \Delta T, 4 + \Delta T]$. The results for different values of ΔT are depicted in Figure 4 for both the UA policy and the SQ policy.

Figure 4a shows that, as expected, the average job allocation rate is constant in the UA policy, even when T has a high variance. This is due to the UA policy's wheel structure that effectively decouples job allocation from worker arrivals. The average job allocation rate under the SQ policy

has a slightly decreasing trend. Furthermore, Figure 4b depicts that the standard deviation of the job allocation rate in the UA policy is significantly lower than in the UA policy. Although in both policies the standard deviation of the allocation rate is increasing with the variance of T, the standard deviation of the UA policy is significantly lower than the SQ policy.

8. Conclusions

We developed a novel modeling framework for the operations of an online labor platform that faces both supply and demand uncertainty and balances profits, worker welfare and customer satisfaction. We suggest a class of analytically-tractable allocation policies called Uniform Allocation (UA), that guarantees near-predictable worker profits. Policies in this class allocate jobs to workers in a rotating manner and allow us to explicitly describe the relationship between platform profits, job wait times and variability of worker revenues.

Through a large-capacity system analysis of UA policies, we obtain the optimal values for the fees the platform charges workers and the job delivery control parameters. These values correspond to the solution of the fluid/deterministic problem corrected by square root terms, proving that the corrected fluid values together with the uniform allocation mechanism are asymptotically optimal. We validate the performance of the UA policy and compare it with a state-dependent policy through a discrete event simulation.

Our work opens multiple promising research avenues. First, a transient analysis of the workers' queue in the platform could lead to a more precise understanding of job wait times. Extending our results to non-homogeneous jobs is both relevant and interesting and would widen the applicability of our model. We assume that traffic follows a stationary Poisson process, which is a valid assumption at an aggregated level. However, exploring time non-homogeneous processes is more realistic and would lead to insight into platforms where supply and demand vary throughout the day. Finally, empirical research on workers, in particular on understanding what are the drivers of their engagement time, as well as customer behavior on online labor platforms could lead to more realistic models of worker welfare and customer satisfaction.

References

Adan, Ivo, Gideon Weiss. 2012. Exact FCFS Matching Rates for Two Infinite Multitype Sequences. *Operations Research* **60**(2) 475–489.

Afèche, Philipp, Zhe Liu, Costis Maglaras. 2017. Ride-hailing networks with strategic drivers : The impact of platform control capabilities on performance.

Asmussen, S. 2003. *Applied Probability and Queues*. Springer-Verlag (Second Edition), New York.

Ata, Barış, Tava Lennon Olsen. 2009. Near-Optimal Dynamic Lead-Time Quotation and Scheduling Under Convex-Concave Customer Delay Costs. *Operations Research* **57**(3) 753–768.

Banerjee, Siddhartha, Ramesh Johari, Carlos Riquelme. 2015. Pricing in Ride-Sharing Platforms: A Queueing-Theoretic Approach. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. EC '15, ACM, New York, NY, USA, 639–639.

Benjaafar, S., G. Kong, X. Li, C. Courcoubetis. 2018. Peer-to-Peer Product Sharing: Implications for Ownership, Usage, and Social Welfare in the Sharing Economy. *Management Science* .

Berg, J., M. Furrer, E. Harmon, U. Rani, M. Silberman. 2018. Digital labour platforms and the future of work: Towards decent work in the online world. Report, International Labour Organization, Geneva.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, Viral B Shah. 2017. Julia: A fresh approach to numerical computing. *SIAM review* **59**(1) 65–98.

Bimpikis, Kostas, Ozan Candogan, Daniela Saban. 2016. Spatial Pricing in Ride-Sharing Networks. SSRN Scholarly Paper ID 2868080, Social Science Research Network, Rochester, NY.

Cachon, Gérard P., Kaitlin M. Daniels, Ruben Lobel. 2017. The Role of Surge Pricing on a Service Platform with Self-Scheduling Capacity. *Manufacturing & Service Operations Management* **19**(3) 368–384.

Caldentey, R., E. Kaplan, G. Weiss. 2009. FCFS Infinite Bipartite Matching of Servers and Customers. *Advances in Applied Probability* **41**(3) 695–730.

Chen, M Keith, Judith A Chevalier, Peter E Rossi, Emily Oehlsen. 2017. The value of flexible work: Evidence from uber drivers. Tech. rep., National Bureau of Economic Research.

Daniels, Kaitlin, Michal Grinstein-Weiss. 2018. The impact of the gig-economy on financial hardship among low-income families. *Available at SSRN 3293988* .

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Garg, Nikhil, Hamid Nazerzadeh. 2019. Driver surge pricing. *arXiv preprint arXiv:1905.07544* .

Gray, Mary L, Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books.

Gurvich, Itai, Martin Lariviere, Antonio Moreno. 2016. Operations in the On-Demand Economy: Staffing Services with Self-Scheduling Capacity. SSRN Scholarly Paper ID 2336514, Social Science Research Network, Rochester, NY.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.

Hu, Ming, Yun Zhou. 2016. Dynamic Type Matching. SSRN Scholarly Paper ID 2592622, Social Science Research Network, Rochester, NY.

Hu, Ming, Yun Zhou. 2017. Price, Wage and Fixed Commission in On-Demand Matching. SSRN Scholarly Paper ID 2949513, Social Science Research Network, Rochester, NY.

King, Peter L., Jennifer S. King, Jennifer S. King. 2016. *The Product Wheel Handbook : Creating Balanced Flow in High-Mix Process Operations*. Productivity Press.

Kingman, J.F.C. 1965. The heavy traffic approximation in the theory of queues. *Proc. Symp. on Congestion Theort* .

Kässi, Otto, Vili Lehdonvirta. 2018. Online labour index: Measuring the online gig economy for policy and research. *Technological Forecasting and Social Change* .

Maglaras, C., J. Meissner. 2006. Dynamic Pricing Strategies for Multiproduct Revenue Management Problems. *Manufacturing & Service Operations Management* **8**(2) 136–148.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53** 242–262.

Moreno, Antonio, Christian Terwiesch. 2014. Doing Business with Strangers: Reputation in Online Service Marketplaces. *Information Systems Research* **25**(4) 865–886.

Ozkan, E., A. Ward. 2017. Dynamic Matching for Real-Time Ridesharing. Working paper.

Rochet, J., J. Tirole. 2006. Two-Sided Markets: A Progress Report. *The RAND Journal of Econ.* **37**(3) 645–667.

Savin, S. V., M. A. Cohen, N. Gans, Z. Katalan. 2005. Capacity management in rental businesses with two customer bases. *Oper. Res.* **53** 617–631.

Taylor, Terry A. 2018. On-Demand Service Platforms. *Manufacturing & Service Operations Management* **20**(4) 704–720.

Vakharia, Donna, Matthew Lease. 2013. Beyond AMT: An Analysis of Crowd Work Platforms. *arXiv:1310.1672 [cs]* ArXiv: 1310.1672.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Sci.* **38**(5) 708–723.

Wilson, Lonnie. 2015. *How To Implement Lean Manufacturing, Second Edition*. 2nd ed. McGraw-Hill Education, New York, N.Y.

Çelik, Sabri, Costis Maglaras. 2008. Dynamic Pricing and Lead-Time Quotation for a Multiclass Make-to-Order Queue. *Management Science* **54**(6) 1132–1146.

APPENDIX A: Main Proofs

A1. Proof of Proposition 1. The key feature of this setting is the constant engagement period T, required by all workers. Despite the uncertainty in the arrival of jobs, the fact that workers spend all T in the system, workers will leave the platform in the order they arrived. We rank the slots from 1 to κ and, when multiple slots are free, we assign workers to the lowest ranking slot. Because of that, we can ex ante assign all the workers that will be approaching the platform to a specific slot. Basically, if we rank workers arriving to the platform $1, 2, ...$ and consider the sequence of workers $(i + m\,\kappa : m \geq 0)$, with $1 \leq i \leq \kappa$, then this sequence of workers will be assigned to slot i. We can then tell, on arrival, to which slot (among the κ) the worker will be allocated to. Therefore, the system can be collapsed to κ single server queues where each queue has the special feature, where a worker will spend exactly T in the system, whether being passive or active. We focus on one queue. Note that the inter-arrival time between two workers sharing the same slot is $V_{j+1} = \sum_{l=j+1}^{j+\kappa} v_l \stackrel{D}{=} \sum_{l=1}^{\kappa} v_l$. Lindley's recursion that tracks waiting time in a single server queue implies that the delay at one slot of worker $n+1$ is given with respect to the delay of the worker ahead of her on that same slot, $\tau_{n+1} = [\tau_n + U_n - V_{n+1}]^+$, where U_n is the time the n^{th} worker on that slot spent being active. We conclude that $\tau_{n+1} = [T - V_{n+1}]^+$. By letting n go to infinity we obtain the result.

Finally, note that $\tau_{n+1} \leq T$ and hence, each worker will spend almost surely a positive amount of time being active. Note that this result takes advantage of the fact that T is constant for all workers. If T is random then this is not necessarily true and in this case, there is a positive probability that a worker will leave the system before becoming active.

Finally, we prove ii.). We clearly have that $[T - v_1]^+ \leq [T - \sum_{j=1}^{\kappa} v_i]^+$ *a.s.*. Moreover, by the Strong Law of Large Numbers (SLLN), $\sum_{j=1}^{\kappa} v_i \to +\infty$ *a.s.*, which implies that $\tau(\kappa) \to 0$ *a.s.* □

One can also obtain a closed form expression of the delay function, $\bar{\tau}$.

Corollary 2 *Under a UA mechanism, we have that*

$$\bar{\tau}(\lambda, \kappa) = Te^{-\lambda T} \sum_{j=\kappa}^{\infty} \left(1 - \frac{\kappa}{j+1}\right) \frac{(\lambda T)^j}{j!}. \tag{a1}$$

A2. Proof of Corollary 2. We let $N(T) = \max\{j : A_j \leq T\}$, where A_j is the time of the j^{th} arrival of a worker. Observe that $N(T)$ is a Poisson random variable with rate λT. Hence, for the values of $N(T)$ below κ, $[T - A_\kappa]^+ = 0$ and so,

$$\bar{\tau}(\kappa) = \sum_{j=0}^{\infty} \mathbb{E}\Big[[T - A_\kappa]^+ | N(T) = j\Big]\, \mathbb{P}(N(T) = j) = \sum_{j=\kappa}^{\infty} \mathbb{E}\big[(T - A_\kappa) | N(T) = j\big]\, \mathbb{P}(N(T) = j).$$

Furthermore, we recall that conditioned on $N(T) = j$, the random variables, $\{A_1, A_2, ..., A_j\}$ are distributed as j i.i.d. uniformly distributed random variables on $(0, T)$ and so A_κ is the κ order statistics which is known to be beta distributed with parameters $(\kappa, j + 1 - \kappa)$. Hence, $\mathbb{E}A_\kappa = T/(j+1) \cdot \kappa$, which proves the result. □

A3. Proof of Lemma 1

$$\begin{aligned}
\mathbf{Var}(\Theta(T-\tau)) &= \mathbb{E}\,\mathbf{Var}(\Theta(T-\tau)|\tau) + \mathbf{Var}\,\mathbb{E}[\Theta(T-\tau)|\tau] \\
&= \mathbb{E}\,[\frac{\mu\,(T-\tau)}{\kappa^2} + \mathbf{Var}[\frac{\mu\,(T-\tau)}{\kappa}] \\
&= \frac{\mu}{\kappa^2}(T-\bar{\tau}) + \frac{\mu^2}{\kappa^2}\,\mathbf{Var}(\tau) \\
&= \frac{\mu}{\kappa^2}(N\kappa/\mu) + \frac{\mu^2}{\kappa^2}\,\mathbf{Var}(\tau) \\
&= \frac{N}{\kappa} + \frac{\mu^2}{\kappa^2}\,\mathbf{Var}(\tau)
\end{aligned}$$

A4. Proof of Lemma 2, Lemma 3 and Proposition 4 *i.*). We recall that each worker is represented as a single server queue that is assumed to reach stationarity instantly. Moreover, we know that the processing time is exponentially distributed with rate γ while the arrival is a renewal process with inter-arrival times that are the sum of κ r.v.'s that are i.i.d. with an exponential distribution with rate μ. Hence, the worker's single server queue is a $G/M/1$ system with this special arrival process. The utilization of this system is equal to $\mu/(\kappa\gamma)$ which is assumed to be strictly less than one.

It is known (see, Asmussen (2003)) that for a $G/M/1$ queue the number of customers in the system follows a geometric distribution and hence, the so-called sojourn time (waiting time plus the service time) is exponentially distributed. It can be shown that the rate of this distribution is equal to $\frac{1}{\gamma\,(1-\nu)}$, where ν is solution to the equation:

$$\nu = \mathbb{E}_X \exp(-\gamma(1-\nu)\,X), \tag{a2}$$

where X is the interarrival times: $X = \sum_{i=1}^{\kappa} u_i$ and $u_i \sim \exp(\mu)$. Hence,

$$W^{\text{UA}} = \frac{1}{\gamma(1-\nu)}.$$

It can also be shown that the equation defining ν admits a unique solution in $(0, 1)$ if and only if the utilization of the queueing system is strictly less than one, which is the case here.

Recalling that the mgf of an exponential r.v. is given by $\mathbb{E}\exp(\theta\, u_1) = \frac{\mu}{\mu-\theta}$. We then write that

$$\begin{aligned}\log\left(\mathbb{E}_X \exp(-\gamma(1-\nu)\, X)\right) &= \log\left(\mathbb{E}\exp\Big(\sum_{i=1}^{\kappa}(-\gamma(1-\nu)\, u_i)\Big)\right)\\ &= \log \Pi_{i=1}^{\kappa}\mathbb{E}\exp\big(-\gamma(1-\nu)\, u_1\big)\\ &= \kappa \log\big(\frac{\mu}{\mu+\gamma(1-\nu)}\big)\\ &= \kappa\, \log\, \mu - \kappa\, \log\big(\mu(1+\gamma(1-\nu)/\mu)\big)\\ &= -\kappa\, \log(1+\gamma(1-\nu)/\mu\big).\end{aligned}$$

As we scale the system, we have that for each element of the sequence of problems, ν^n is solution to

$$\log \nu^n = -\kappa^n\, \big(\frac{\gamma}{\mu^n}\,(1-\nu^n) + O(\frac{1}{\mu^{n^2}})\big),$$

as $n\to\infty$. In the above, we took advantage of the fact that at the limit the utilization is strictly less than one and hence $\nu^n \to \nu^0 \in (0,1)$ as $n\to\infty$, with ν^0 solution to

$$\frac{\log \nu}{1-\nu} = -\frac{\kappa^0}{\mu}\gamma = -\frac{T\gamma}{N}.$$

Equivalently, we can write that

$$\frac{\log \nu^n}{1-\nu^n} = -\frac{\kappa^n}{\mu^n}\,\gamma + O(1/n) = -\frac{T\gamma}{N}(1-\frac{\bar\tau}{T\sqrt{n}}) + O(\frac{1}{n}),$$

as $n\to\infty$.

A similar analysis applies in the fluid case where the interarrivals are constant equal to κ^0/μ, so that if $\kappa^0 = \mu\, T/N$ then ν^0 will be the solution to

$$\frac{\log \nu}{1-\nu} = -\kappa^0\, \frac{\gamma}{\mu} = -\frac{T\gamma}{N}.$$

We denote by $\Gamma(x) = -x\, \log\big(\frac{x-1}{x}\big)$ is a decreasing function on $(1,\infty)$ that admits an inverse Γ^{-1} that is also decreasing on $(1,\infty)$. Note that $\Gamma(\frac{1}{1-\nu}) = -\frac{\log\nu}{1-\nu}$. We then write that

$$\begin{aligned}\frac{1}{1-\nu^0} - \frac{1}{1-\nu^n} &= \Gamma^{-1}(\frac{T\gamma}{N}) - \Gamma^{-1}\left(\frac{T\gamma}{N}(1-\frac{\bar\tau}{T\sqrt{n}}) + O(\frac{1}{n})\right)\\ &= \frac{\gamma\bar\tau}{N\sqrt{n}}\,\Gamma^{-1\prime}(\frac{T\gamma}{N}) + O(1/n)\end{aligned}$$

as $n\to\infty$. We let $\beta = -\Gamma^{-1\prime}(\frac{T\gamma}{N}) > 0$. Recalling the formulation of the waiting time of such $\cdot/M/1$, we conclude from the above that:

$$\begin{aligned}C^{\mathrm{UA},n} - C^0 &= cN(W^{\mathrm{UA},n} - W^0)\\ &= cN\left(\frac{1}{\gamma(1-\nu^n)} - \frac{1}{\gamma(1-\nu^0)}\right)\\ &= cN\left(-\frac{\bar\tau}{N\sqrt{n}}\,\Gamma^{-1\prime}(\frac{T\gamma}{N}) + O(1/n)\right)\\ &= c\beta(N)\frac{\bar\tau}{\sqrt{n}} + O(1/n)\end{aligned}$$

as $n \to \infty$. □

A5. Proof of Proposition 2.

We start by showing that the fluid profit is unimodal as long as $H(x) := x\, F_\xi^{-1}(x)$ is convex.

We rewrite the fluid profit as follows:

$$\Pi^0(\lambda, N) = r\,\lambda\, N - W(N)H(\lambda) - c\,W(N) - T\,H(\lambda)$$

We focus on the function

$$\mathcal{G}(\lambda, N) = r\lambda\, N - W(N)H(\lambda).$$

Assume that N_1 and λ_1 define a stationary point for $\mathcal{G}$ so that the two partial derivatives at that point equal zero:

$$rN_1 - W(N_1)H'(\lambda_1) = 0 \quad \text{and} \quad r\lambda_1 - W'(N_1)H(\lambda_1) = 0.$$

Consider any perturbation around (λ_1, N_1). For instance, suppose δ and δ' are positive and consider the difference

$$\begin{aligned}
&\mathcal{G}(\lambda_1 + \delta', N_1 + \delta) - \mathcal{G}(\lambda_1, N_1) \\
&\quad = r(\lambda_1 + \delta')\,(N_1 + \delta) - W(N_1 + \delta)\,H(\lambda + \delta') - \mathcal{G}(\lambda_1, N_1) \\
&\quad = r(\delta'\,N_1 + \delta\,\lambda_1) - \delta'\,W(N_1)\,H'(\lambda_1) - \delta\,W'(N_1)\,H(\lambda_1) + A \\
&\quad = \delta'(rN_1 - W(N_1)H'(\lambda_1)) + \delta(r\lambda_1 - W'(N_1)H(\lambda_1)) + A \\
&\quad = A,
\end{aligned}$$

where

$$A = r\delta\delta' - \delta^2\, W''(N_1)H(\lambda) - \delta'^2 W(N_1)H''(\lambda)$$

We could have considered any perturbation of (λ_1, N_1) and we would obtain the same sign of the inequality. This shows that the stationary point must be a local maximum and hence by continuity the stationary point is unique.

Given that the function is unimodal and if its unconstrained maximizer is such that $\bar{\lambda}\,\bar{N}/\mu > 1$, then it must be that the optimizer of Π^0 under the constraint $\rho \leq 1$ is such that ρ is exactly one.

□

A6. Proof of Theorem 1 and Proposition 3

We will prove Theorem 1 and Proposition 3 at the same time. We start with a lemma.

Lemma 4 *Let X be a normal random variable with mean η and standard deviation σ. The expected value of the truncated normal is given by $\mathbb{E}[X]^+ = \sigma\Psi(-\eta/\sigma)$, where the function Ψ is defined for all x, as $\Psi(x) = \phi(x) - x\bar{\Phi}(x)$ (see Section 5). Furthermore, Ψ is decreasing on $\mathbb{R}$ and for all $x \in \mathbb{R}$, $\Psi(-x) > x$ with $\Psi(-x)/x \to 1$, as $x \to +\infty$.*

By definition $\Psi(x) = \phi(x) - x\bar{\Phi}(x) \geq -x$. $\Psi(x)/x = \phi(x)/x - \bar{\Phi}(x) \to -1$ as $x \to -\infty$.

We move to the proof of Proposition 3.

• We recall here an important result that is a consequence of Bolzano-Weierstrass theorem: if a bounded sequence of reals has the property that every subsequence that is convergent has the same limit, then the whole sequence is itself convergent to that same limit.

• Let $\mu^n = n\mu$, and $\Lambda^n = n\,\Lambda$, while $T^n = T$. We fix a value of N that the platform is targeting. For a given (λ^n, κ^n, N) there exists a price p^n that solves the equilibrium equation. Most of the analysis will be done for any feasible p_n but eventually the objective is to maximize the profit and the optimal price will be denoted by p^n_*.

• We start by noting that both sequences: $(\frac{\kappa^n}{n} : n \geq 0)$ and $(\frac{\lambda^n}{n} : n \geq 0)$ are bounded from above. Indeed, notice that any solution to the fulfillment constraint requires that $T - \frac{N\kappa^n}{n\mu} \geq 0$ and hence, we conclude that $\frac{\kappa^n}{n} \leq \frac{\mu T}{N}$. Furthermore, we only consider prices such that the utilization will be bounded by one which requires that $\frac{\lambda^n}{n} \leq \frac{\mu}{N}$. Finally, we will restrict ourselves without loss of generality to pricing sequences such that λ^n/n are bounded from below. From the equilibrium equation we can see that this holds as long as $\sup_{n \geq n_0} p^n < r$ for some n_0. This will be easily checked at the end of the proof.

• We want to show that $\sqrt{n}\,(\sum_1^{\kappa^n} v_i - \frac{\kappa^n}{\lambda^n}) \Rightarrow Y \stackrel{d}{=} \sigma^0\, Z$ as $m \to \infty$, where Z is a standard normal random variable and $\sigma_0 > 0$. For that we consider the log-moment generating function of the quantity $\sum_{i=1}^{\kappa^n} v_i - \frac{\kappa^n}{\lambda^n}$. As long as λ^n/n is bounded away from zero, we take advantage of the fact that v_j is an i.i.d. sequence of exponential r.v.'s', and by a simple Taylor expansion we write

$$\log \mathbb{E} \exp \theta(\sum_{i=1}^{\kappa^n} v_i - \frac{\kappa^n}{\lambda^n}) = \theta^2 \frac{\kappa^n}{2\lambda^{n\,2}} + o(n^{-2}), \tag{a3}$$

as $n \to \infty$.

• Consider any converging subsequence of κ^n/n and another converging subsequence of λ^n/n. Let $m = (m^n : n \geq 0)$ be a common subsequence. Form the bounded subsequence $m_n \kappa^{m_n}/\lambda^{m_n\,2}$; it must converge. We denote by l_m its finite limit. For clarity of exposition we index from now on the subsequence by m instead of m_n, i.e., $m\kappa^m/\lambda^{m\,2} \to l_m$ as $m \to \infty$. Recalling the expression of the log-moment generating function, we conclude that $\sqrt{m}\,(\sum_1^{\kappa^m} v_i - \frac{\kappa^m}{\lambda^m}) \Rightarrow Y \stackrel{d}{=} \sigma^0_m\, Z$ as $m \to \infty$, where $\sigma^0_m = \sqrt{l_m}$.

- From Proposition 1, the fulfillment constraint can be written as follows:

$$\bar{\tau}^m = \mathbb{E}[T - \sum_{i=1}^{\kappa^m} v_i]^+ = T - \frac{N\kappa^m}{m\,\mu}.$$

Hence,

$$\sqrt{m}\,(T - \frac{N\kappa^m}{m\mu}) = \sqrt{m}\,\mathbb{E}\left[T - \frac{\kappa^m}{\lambda^m} - \left(\sum_{i=1}^{\kappa^m} v_i - \frac{\kappa^m}{\lambda^m}\right)\right]^+ = \mathbb{E}\big[\sqrt{m}\,(T - \frac{\kappa^m}{\lambda^m}) + Y + \varepsilon_m\big]^+, \qquad \text{(a4)}$$

where $\varepsilon_m \to 0$ as $m \to \infty$. Equivalently, we have,

$$0 = \mathbb{E}\max\big\{\sqrt{m}(\frac{N\kappa^m}{m\mu} - \frac{\kappa^m}{\lambda^m}) + Y + \varepsilon_m, -\sqrt{m}(T - \frac{N\kappa^m}{m\mu})\big\}.$$

- Consider a first regime made of subsequences of m (we use now the index j) for which $\sqrt{j}(T - \frac{N\kappa^j}{j\mu})$ is divergent.

— For such subsequences, and for the fulfillment constraint to hold, the first term in the above maximum must go to zero in expected value as j gets large:

$$\sqrt{j}(\frac{N\kappa^j}{j\mu} - \frac{\kappa^j}{\lambda^j}) \to 0,$$

as $j \to \infty$.

— In particular, $\frac{N\kappa^j}{j\mu} - \frac{\kappa^j}{\lambda^j} = \frac{\kappa^j}{\lambda^j}(\rho^j - 1) \to 0$ as $j \to \infty$. This convergence implies that $\rho^j \to 1$ i.e. $\lambda^j/j \to \lambda^0 = \frac{\mu}{N}$ and in turn $\kappa^j/j \to {\lambda^0}^2 l_0$. From the fulfillment constraint we conclude that

$$\bar{\tau}^j = T - \frac{N\,\kappa^j}{j\,\mu} \to T - \frac{N{\lambda^0}^2 l_0}{\mu} \geq 0.$$

- The other possible regime is made of all subsequences for which $\sqrt{m}(T - \frac{N\kappa^m}{\mu^m})$ are bounded.

— Consider in such regime any converging subsequence(indexed by j):

$$\sqrt{j}(T - \frac{N\kappa^j}{j\mu}) \to \bar{\tau},$$

for some non-negative finite $\bar{\tau}$. For that to occur, and as we recall equation a4, we must have

$$\sqrt{j}\,(T - \frac{\kappa^j}{\lambda^j}) \to \eta$$

for some finite η.

— From the first limit, we conclude that $\sqrt{j}\,\tau^j \to \bar{\tau}$, as $j \to \infty$, so that $T - \frac{N\kappa^j}{j\mu} \to 0$. We conclude that $\kappa^j/j \to \kappa^0 = \mu T/N$. From the second limit we conclude: $\lambda^j/j \to \kappa^0/T = \frac{\mu}{N}$.

—Based on Lemma 4 (stated above) we have that

$$\sqrt{j}\,\bar{\tau}^j = \mathbb{E}\big[\sqrt{j}\,(T - \frac{\kappa^j}{\lambda^j}) + Y + \varepsilon_j\big]^+ \to \sigma_0 \Psi(-\eta/\sigma_0),$$

as $j \to \infty$, where ${\sigma^0}^2 = \kappa^0/{\lambda^0}^2$. Note that (for now) η depends on the subsequence n_j. We conclude from the fulfillment constraint (at the limit) that $\sigma^0 \Psi(-\eta/\sigma^0) = \bar{\tau}$. We just showed that,

$$\sqrt{j}\,\bar{\tau}^j = \bar{\tau} + o(1),$$

as $j \to \infty$.

- From both possible regimes, we conclude that all converging subsequences of λ^n/n converge respectively to λ^0 as $n \to \infty$.
- Hence, in both regimes, the revenue side of the profit is maximized at the limit (with $\lambda^n \to \lambda^0$ or equivalently $\rho^n \to 1$ as $n \to \infty$). However, in the first regime: $\sqrt{j}\,\tau^j \to \infty$, while for the second regime this limit is finite. We thus conclude that the second regime always outperforms the first one i.e. no matter the subsequence, as j gets larger the solution to the profit maximization will satisfy the condition of the the second regime.
- We now inject in the fulfillment constraint, the formulation of $\bar{\tau}^j$ and solve for κ^j. We obtain that $\kappa^j = (T - \frac{\bar{\tau}}{\sqrt{j}} + o(1/\sqrt{j}))\frac{n\,\mu}{N} = \kappa^0 j - \frac{\mu\,\bar{\tau}}{N}\sqrt{j} + o(\sqrt{j})$.
- In turns, we inject the expression of κ^j in the term $(\sqrt{j}(T - \kappa^j/\lambda^j)$ and get that

$$\sqrt{j}\,(T - \kappa^j/\lambda^j) = \sqrt{j}\,\left(T - \frac{\mu\,T/N\,j - \frac{\mu\,\bar{\tau}}{N}\sqrt{j} + o(\sqrt{j})}{\lambda^j}\right) \tag{a5}$$

$$= \sqrt{j}\,T(1 - (\rho^j)^{-1}) + (\rho^j)^{-1}\,\bar{\tau} + o(1) \tag{a6}$$

$$= \sqrt{j}\,T(\rho^j - 1)/\rho^j + \frac{\bar{\tau}}{\rho^j} + o(1) \tag{a7}$$

as $j \to \infty$. Given that $\rho^j \to 1$ as $j \to \infty$, this implies that $\sqrt{j}\,(1 - \rho^j) \to d$, for some $d \geq 0$ so that

$$\eta = -d\,T + \bar{\tau}.$$

- We write $\lambda^j = \lambda^0 j - l^j$. From the limiting result of the utilization, we imply that $(l^j\,N)/(\sqrt{j}\mu) \to d$, from which we conclude that $\lambda^j = \lambda^0\,j - \frac{\mu d}{N}\sqrt{j} + o(\sqrt{j})$.
- Note that if $d = 0$, then $\eta = \bar{\tau} = \sigma\,\Psi(-\eta/\sigma)$ and that equation does not have any solution in η (Lemma 1). Hence, $d = (\bar{\tau} - \eta)/T > 0$, and, $\lambda^j = \lambda^0\,j - \lambda^0/T\,(\bar{\tau} - \eta)\,\sqrt{j} + o(\sqrt{j})$.
- The pricing policy that guarantees this arrival can be implied from a Taylor expansion of $\lambda^j(\cdot)$ in the neighborhood of $p^0 := \lambda^{-1}(\lambda^0) = r - \frac{T + W^{\mathrm{UA},0}(N)}{N} F_\xi^{-1}(\frac{\lambda^0}{\Lambda})$. We write $\lambda^j(p^j) = \lambda^0 j + (p^j - p^0)\lambda'^0 j + o((p^j - p^0)j)$, where $\lambda^{0\prime} = \lambda'(p^0)$, the first derivative of λ at p^0, so that $\lambda^{0\prime} = -\frac{\Lambda\,N}{T + W^{\mathrm{UA},0}(N)} f_\xi\left((r-p)\frac{N}{T + W^{\mathrm{UA},0}(N)}\right)$

- By comparing the two expressions of λ^j as j is large, we conclude that

$$p^j = p^0 - \frac{\lambda^0\,(\bar{\tau}-\eta)}{\lambda^{0\prime}\,T}\frac{1}{\sqrt{j}} + o(1/\sqrt{j}).$$

Recall that $p^0 < r$ and hence, for j large enough $p^j < r$.

- The entire policy is constructed at this point. We still have a free parameter η to determine (which for now depends on the subsequence indexed by j). We recall that the profit obtained in the deterministic setting is $\Pi^{0,j} = \lambda^0\,N^0\,j(p^0 - C^{\mathrm{UA},0^j})$ which is an upper bound of the the profit rate in the stochastic case. The parameter η will be selected in order to maximize that ratio for large j. We recall that $C^{\mathrm{UA},0^j} = O(1)$ and set $c^0 = \frac{T}{N}\beta(N)\,c$. We write that

$$\begin{aligned}
\Pi^j(\lambda^j,\kappa^j,N) &= \lambda^j\,p^j(\lambda^j,\kappa^j,N)\,N - \lambda^j\,\Sigma^{\mathrm{UA}^j}(\lambda^j,\kappa,N) - \lambda^j\,C^{\mathrm{UA}^j}(\kappa,N)\\
&= \lambda^j\,\left(p^j\,N - b\,N/T\,\frac{\sigma_\tau}{\sqrt{j}} + o(1/\sqrt{j}) - C^{\mathrm{UA},0^j} - c\,\beta(N)\frac{\bar{\tau}}{\sqrt{j}} + o(1/\sqrt{j})\right)\\
&= \big(\lambda^0 j - \lambda^0/T\,(\bar{\tau}-\eta)\sqrt{j} + o(\sqrt{j})\big)\cdot\left(\big(p^0 - \frac{\lambda^0\,(\bar{\tau}-\eta)}{\lambda^{0\prime}\,T}\frac{1}{\sqrt{j}} + o(1/\sqrt{j})\big)\,N - C^{\mathrm{UA},0^j}\right)\\
&\qquad - N/T\big(\lambda^0 j - \lambda^0/T\,(\bar{\tau}-\eta)\sqrt{j} + o(\sqrt{n})\big)(c^0\,\bar{\tau}/\sqrt{j} + b\,\sigma_\tau/\sqrt{j} + o(1/\sqrt{j}))\\
&= \lambda^0\,j\big(p^0\,N - C^{\mathrm{UA},0}\big) - p^0\lambda^0 N\,(\bar{\tau}-\eta)/T\,\sqrt{j} - \frac{\lambda^{0^2}\,(\bar{\tau}-\eta)N}{\lambda^{0\prime}\,T}\sqrt{j} - c^0 N/T\lambda^0\bar{\tau}\sqrt{j}\\
&\qquad - bN/T\lambda^0\,\sigma_\tau\sqrt{j} + O(1)\\
&= \lambda^0\,j\big(p^0\,N - C^{\mathrm{UA},0}\big) - \lambda^0 p^0 N/T\big[\bar{\tau}\,(1+\lambda^0/(\lambda^{0\prime}p^0) + c^0/(p^0)) + b/p^0\,\sigma_\tau - \eta(1+\lambda^0/(\lambda^{0\prime}p^0))\big]\,\sqrt{j}\\
&\qquad + O(1)\\
&= \lambda^0\,j\big(p^0\,N - C^{\mathrm{UA},0}\big) - \lambda^0 p^0 N/T\big[\bar{\tau}\,(1+1/e^0 + c^0/p^0) + b/p^0\,\sigma_\tau - \eta(1+1/e^0)\big]\,\sqrt{j} + O(1).
\end{aligned}$$

Hence, at $N = N^0$, we have that

$$\frac{\Pi^j}{\Pi^{0,j}} = 1 - \frac{\mu p^0}{\mu p^0 - \lambda^0 C^{\mathrm{UA},0}}\frac{\xi(\eta)}{T\sqrt{j}} + o(1/\sqrt{j}),$$

where

$$\xi(\eta) = \big[\bar{\tau}\,(1+1/e^0 + c^0/p^0) + b/p^0\,\sigma_\tau - \eta(1+1/e^0)\big]\,.$$

We pick η, so as to minimize $\xi(\eta)$. The function $\xi(\cdot)$ is convex. We take the derivative of ξ with respect to η and recall that $\bar{\tau}'(\eta) = \bar{\Phi}(-\eta/\sigma)$ and $\sigma_\tau(\eta) = \sigma^2\,\chi(-\eta/\sigma)$. We show that there exists a unique η^* that minimizes $\xi(\cdot)$. If $b = 0$, then $\eta_c^* = -\sigma\bar{\Phi}^{-1}\left(\left(1 + \frac{c^0}{p^0(1+1/e^0)}\right)^{-1}\right)$ as long as $e^0 < -1$.

- The above also proves that the constant η is unique independent of the subsequence, which also means that all the subsequences of λ^n and κ^n are asymptotically the same, proving the result.

□

A7. Proof of Proposition 4 *ii.*).

Recall from Lemma 1 that

$$\mathrm{Var}(\hat{\Theta}(T-\tau)) = \frac{\mu}{\kappa^2}\left(\mathrm{Var}(\tau) + (T-\bar{\tau})^2\right) + \frac{\mu^2}{\kappa^2}\mathrm{Var}(\tau)$$

By scaling the above equality and recalling from the proof of Theorem 1 that $\sqrt{n}\,\tau^n \Rightarrow X^+$ where $X \sim \mathcal{N}(\eta, \sigma)$, we obtain that the first term must go to zero, while the second is of the form

$$\frac{(\mu^n)^2}{n\,(\kappa^n)^2}\mathrm{Var}(\sqrt{n}\,\tau^n) \to \frac{\mu^2}{(\kappa^0(N))^2}\,\sigma_\tau^2$$

as $n \to \infty$ □

APPENDIX B: Additional Proofs for Target-N Engagements

B1. N-Engagements

We denote by N-Engagement the setting where the engagement of a worker is driven by a target of N jobs. We introduce two results related to N-Engagements.

Proposition 5

i.) Under an N-Engagement, the delay is given by

$$\bar{\tau}(\lambda,\kappa) = \mathbb{E}\max_{n\geq 0}\sum_{i=1}^{n} X_i(\lambda,\kappa),$$

where the sequence $(X_1, X_2, ...)$ *is i.i.d. with* $X_1 \stackrel{d}{=} \sum_{i=1}^{N\kappa} u_i - \sum_{i=1}^{\kappa} v_i$, *and* u_i*'s and* v_i*'s are the interarrival times of the jobs and the workers, respectively.*

ii.) Moreover,

$$\textit{for all}\ \ \kappa \geq 1, \quad \bar{\tau}(\kappa) \leq \bar{\tau}(1) \qquad \textit{and} \qquad \bar{\tau}(\kappa) \to 0, \quad \textit{as}\ \ \kappa \to \infty.$$

Proof. The key feature of this setting is the constant number of jobs, N, required by all workers. Despite the uncertainty in the arrival of jobs, such uncertainty does not alter the order of the workers leaving the system (after having their met their fulfillment constraint). This order is the same than the one they had when they initially approached the publisher. We rank the slots from 1 to κ and, when multiple slots are free, we assign workers to the lowest ranking slot. We can then tell, at arrival, on which slot (among the κ available) the ad will be displayed. Therefore, the slots dynamics can be decoupled each having its arrival process.

Let U_i be the time spent by i^{th} worker active. The sequence $U = (U_i : i \geq 1)$ is stationary. Every κ jobs is directed to the same worker, and every worker is requesting N jobs. Thus, $U_1 \stackrel{D}{=} \sum_{j=1}^{N\kappa} u_j$ where the u_j's are the interarrival times between jobs. Similarly, let $V_{j+1} = \sum_{l=j+1}^{j+\kappa} v_l \stackrel{D}{=} \sum_{l=1}^{\kappa} v_l$, where, v_l's are the interarrival times between workers. Similarly to the dynamics of a single server queue, we can track the delay of each worker. Assume that the n^{th} was assigned a certain slot (among the κ) then the next worker that will be assigned the same slot is the $(n+\kappa)^{th}$ worker received. The arrival time between two consecutive workers sharing the same slot is $\sum_{l=n+1}^{n+\kappa} v_l = V_{n+1}$. The formulation of the engagement delay of a worker follows a Lindley's type recursion $W_{n+\kappa} = [W_n + U_n - V_{n+1}]^+$. Notice here that W_n is independent of U_n and V_{n+1}. Unfolding this recurrent equation leads to $W_n \stackrel{D}{=} \max_{0\leq m\leq n} S_m(\kappa)$ with $S_m(\kappa) = \sum_{j=1}^{m} X_j$ and $X_j = U_{j-\kappa} - V_{j-\kappa+1}$. Observe that X_1 is the difference between two gamma distributed random variable (and not the difference between two exponentially distributed r.v.). This Lindley relationship implies that the stationary distribution of the delay exists and is finite almost surely. Furthermore, it is equal in

distribution to an infinite horizon maximum of a random walk $W_n \Rightarrow M(\kappa) = \max_{n\geq 0} S_n(\kappa)$, as $n \to \infty$. Of course, $W_{sn+1}, W_{sn+2}, ..W_{sn+}$ are dependent random variables as their associated worker is fulfilled with (at least partially) the same jobs. However, all these variables converge weakly to the same random variable M and hence, W_n as well. This single server queue type-relationship implies that when both u_i's and v_i's are exponentially distributed, then the delay function is equal in distribution to the waiting of a single server queue with interarrival times and service times distributed respectively as gamma random variables.

As for *ii.*) In this case, $\tau(\kappa) = \max_{m\geq 0} S_m(\kappa) \stackrel{d}{=} \max_{m\kappa\geq 0} S_m(1) \leq \max_{m\geq 0} S_m(1) \stackrel{d}{=} \tau(1)$. By the SLLN, $S_m(\kappa) \to -\infty$ *a.s.* and hence, $\tau(\kappa) \to 0$ *a.s.*. □

Proposition 6 *Consider the setting of an N-engagement. Suppose that the input stream of advertisers follows a Poisson process and both demand and supply are scaled as suggested in Section 5. Assume that* $\lambda^0 \leq \bar{\lambda}$ *and* $\lambda^{0\prime}$ *exists and is finite such that* $e^0 > 1$*. Then, the solution of the optimization problem* (λ^n, κ^n) *is such that*

i.) $\lambda^n = \lambda^0 n - \lambda^0 \frac{\eta^N}{T}\sqrt{n} + o(\sqrt{n})$

ii.) $\kappa^n = \kappa^0 n - \kappa^0 \frac{\bar{\tau}^N}{T}\sqrt{n} + o(\sqrt{n})$

iii.) $\rho^n = 1 - \eta^N/(T\sqrt{n}) + o(1/\sqrt{n})$

iv.) $\bar{\tau}^{N,n}(\lambda^n, \kappa^n) = \bar{\tau}^N/\sqrt{n} + o(1/\sqrt{n})$

v.) If the profit obtained in the deterministic setting is $\Pi^{0,n}$ *then, the ratio* $\Pi^n/\Pi^{0,n}$ *is of the form* $\frac{\Pi^n}{\Pi^{0,n}} = 1 - \beta(\eta)/\sqrt{n} + o(1/\sqrt{n})$,

where, $\bar{\tau}^N = \mathbb{E}\max_{r\geq 0} S_r$*, and* $(S_r : r \geq 0)$ *is a random walk with normally distributed increments with mean* η^N *and standard deviation* $\sigma = (\frac{\kappa^0 N}{\mu^{0^2}} + \frac{\kappa^0}{\lambda^{0^2}})^{1/2}$*;* η^N *is selected so that* $\beta(\eta)$ *is minimized.*

We do have approximations of $\bar{\tau}^N$. One of them, $\bar{\tau}^N \approx \frac{\sigma^2}{2\eta^N}$ is given by Kingman (1965). In teh case where $b \equiv 0$, if we replace $\bar{\tau}^N$ by this approximation, the optimal value of β is given by $\beta^* = \eta^N/T(1 + 1/e^0) + c^0 N/T/(p^0 N)\sigma^2/(2\eta^N)$, and $\eta^N = \sigma\sqrt{\frac{c}{2p^0(1+1/e^0)}}$, when again $e^0 < -1$. The proof follows the same approach as the proof of the T-engagement model. We will only describe the parts that are different

Proof. For the sake of the proof, we drop the index N. We recall that

$$W^n \stackrel{d}{=} \max_{r\geq 0} S_r^n(\kappa^n),$$

where $S_r^n(\kappa^n) = \sum_{i=1}^r Y_i^n$ where, $Y_i^n \stackrel{d}{=} Y_1^n \stackrel{d}{=} \sum_{j=1}^{N\kappa^n} u_j^n - \sum_{j=1}^{1\kappa^n} v_j^n$ with $\mathbb{E}Y_1^n = \kappa^n\,(N/\mu^n - 1/\lambda^n) \leq 0$.

The sequence (λ^n, κ^n) is formed, for every $n \geq 1$, as the solution to the optimization problem (P^n). From the fulfillment constraint we have that $T - N\kappa^n/\mu^n \geq 0$ and hence the sequence $\kappa^n/n \leq$

$\kappa^0 := \mu T/N$. Moreover, the utilization being smaller than one implies that the sequence $\lambda^n/n \leq \lambda^0 := 1\mu/N$. Finally, the sequence κ^n/λ^n is also bounded as λ^n is assumed to be away from zero.

Consider any subsequence κ^m/m that converges to $l < \infty$. The finiteness of such limit l implies that $\kappa^m(N/\mu^{m2} + 1/\lambda^{m2}) \to 0$ as $m \to \infty$. The inter-arrivals of workers and jobs are both exponentially distributed, we conclude that the log-moment generating function of the random variable Y_1^n is given by

$$\log \mathbb{E} \exp \theta Y_1^n = \theta \kappa^n \left(N/\mu^n - 1/\lambda^n\right) + \theta^2 \kappa^n (N/\mu^{n2} + 1/\lambda^{n2}) + O(n^{-2}). \tag{a1}$$

The first term $\kappa^m \left(N/\mu^m - 1/\lambda^m\right) = \kappa^m s/\lambda^m \left(\rho^m - 1\right) \leq 0$ and all other terms go to zero with m. We infer that $\limsup_{m\to\infty} Y^m \leq 0$ a.s. The same holds for S_r^m for all r. Therefore, their maximum, $W^m \Rightarrow 0$ as $m \to \infty$. By bounded convergence, $\mathbb{E}W^m = \varpi^m \to 0$, as $m \to \infty$. From the the equality constraint we conclude that $\varpi^m \to T - Nl/\mu$ as $m \to \infty$. This imposes that $l = \kappa_0 := \mu T/N$ and hence the entire sequence κ^n/n converges to κ_0 as $n \to \infty$.

Similarly, consider a subsequence λ^m/m that converges to some finite limit l' as $m \to \infty$. Consider the log-moment generating function with θ replaced by $\theta\sqrt{m}$. The quantity, $m\,\kappa^m(N/\mu^{m2} + 1/\lambda^{m2}) \to N\kappa_0/\mu^2 + \kappa_0/l'^2$ as $m \to \infty$; While $\limsup_{m\to\infty} \sqrt{m}\,\kappa^m \left(N/\mu^m - 1/\lambda^m\right) = \eta \leq 0$ and possibly infinite. Assume that $\eta < \infty$, in this case $\kappa^m \left(N/\mu^m - 1/\lambda^m\right) \to 0$ and thus $N\kappa_0/\mu - \kappa_0/l' = 0$, equivalently $l' = \lambda_0$ and so all subsequences, that lead to some η finite have that $\lambda^m/m \to \lambda_0$. Any subsequence that lead to an η infinite will still have to satisfy $\lambda^m/m \to 0$; otherwise, it will generate lower profits at the limit. In the finite case, $\lim_{m\to\infty} \sqrt{m}\, Y^m = Y$ where Y is a normal random variable with mean η and standard deviation $\sigma_0 = (N\kappa_0/\mu^2 + \kappa_0/\lambda_0^2)^{1/2}$. As for the delay, we claim that $\sqrt{m}\, W^{N,m} \Rightarrow \max_{r\geq 0} S_r$, where $S_r = \sum_{i=1}^r Y_i$ with Y_i's i.i.d. with $Y_1 \stackrel{d}{=} Y$. To prove it, we rely on Theorem 6.1 on page 285 of Asmussen (2003) which only require uniform integrability of $\sqrt{m}\, Y_i^m$, which is guarantied by the fact that $\mathbb{E}m\, Y_1^{m2} \to \sigma_0^2$ as $m \to \infty$. We denote by $\varpi^N = \mathbb{E} \max_{r\geq 0} S_r$ and $\varpi^m = \varpi^N/\sqrt{m} + o(1/\sqrt{m})$ as $m \to \infty$. The rest of the proof follows the exact same steps as in the T-engagement case. The parameter η is uniquely selected by maximizing the ratio of the profit in the stochastic setting with that in the fluid setting. If the subsequence indexed by m was selected so that η is infinite, in this case, $\varpi^N = 0$ and $\sqrt{m}(T - N\kappa^m/\mu^m) \to 0$ as $m \to \infty$ and thus $\kappa^m = \kappa_0 + o(1/\sqrt{m})$, which implies by injecting κ^m in $\sqrt{m}\,\kappa^m \left(N/\mu^m - 1/\lambda^m\right)$ and recalling that the latter converge to $-\infty$ that $\lambda^m/m = \lambda_0 + l^m$ where $\sqrt{m}\, l^m \to -\infty$. Hence, the demand rate grows at a slower rate than the subsequences corresponding to a finite η. □