Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Asymptotically Optimal Control of a Centralized Dynamic Matching Market with General Utilities

Jose H. Blanchet, Martin I. Reiman, Virag Shah, Lawrence M. Wein, Linjia Wu

To cite this article:
Jose H. Blanchet, Martin I. Reiman, Virag Shah, Lawrence M. Wein, Linjia Wu (2022) Asymptotically Optimal Control of a Centralized Dynamic Matching Market with General Utilities. Operations Research 70(6):3355-3370. https://doi.org/10.1287/opre.2021.2186

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 6, November–December 2022, pp. 3355–3370
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Asymptotically Optimal Control of a Centralized Dynamic Matching Market with General Utilities

Jose H. Blanchet,[a] Martin I. Reiman,[b] Virag Shah,[a] Lawrence M. Wein,[c] Linjia Wu[a]

[a] Management Science and Engineering Department, Stanford University, Stanford, California 94305; [b] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027; [c] Graduate School of Business, Stanford University, Stanford, California 94305

Contact: jose.blanchet@stanford.edu, https://orcid.org/0000-0001-5895-0912 (JHB); martyreiman@gmail.com, https://orcid.org/0000-0003-4919-2894 (MIB); virag@stanford.edu, https://orcid.org/0000-0002-3930-0348 (VS); lwein@stanford.edu, https://orcid.org/0000-0001-6125-0220 (LMW); linjiawu@stanford.edu, https://orcid.org/0000-0002-6941-8902 (LW)

Received: January 24, 2020
Revised: December 7, 2020; April 22, 2021
Accepted: July 25, 2021
Published Online in Articles in Advance: January 21, 2022

Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2021.2186

Abstract. We consider a matching market where buyers and sellers arrive according to independent Poisson processes at the same rate and independently abandon the market if not matched after an exponential amount of time with the same mean. In this centralized market, the utility for the system manager from matching any buyer and any seller is a general random variable. We consider a sequence of systems indexed by n where the arrivals in the nth system are sped up by a factor of n. We analyze two families of one-parameter policies: the population threshold policy immediately matches an arriving agent to its best available mate only if the number of mates in the system is above a threshold, and the utility threshold policy matches an arriving agent to its best available mate only if the corresponding utility is above a threshold. Using an asymptotic fluid analysis of the two-dimensional Markov process of buyers and sellers, we show that when the matching utility distribution is light-tailed, the population threshold policy with threshold $\frac{n}{\ln n}$ is asymptotically optimal among all policies that make matches only at agent arrival epochs. In the heavy-tailed case, we characterize the optimal threshold level for both policies. We also study the utility threshold policy in an unbalanced matching market with heavy-tailed matching utilities and find that the buyers and sellers have the same asymptotically optimal utility threshold. To illustrate our theoretical results, we use extreme value theory to derive optimal thresholds when the matching utility distribution is exponential, uniform, Pareto, and correlated Pareto. In general, we find that as the right tail of the matching utility distribution gets heavier, the threshold level of each policy (and hence market thickness) increases, as does the magnitude by which the utility threshold policy outperforms the population threshold policy.

Funding: J. H. Blanchet received financial support from the U.S. National Science Foundation [Grants 1915967, 1820942, and 1838576].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2021.2186.

Keywords: matching markets • queueing asymptotics • regularly varying functions • extreme value theory

1. Introduction

We consider a symmetric centralized dynamic matching market (the asymmetric case is also discussed for heavy-tailed utilities). Two types of agents, which we call buyers and sellers, arrive to the market according to independent Poisson processes with rate λ, and each agent abandons (i.e., exits) the market after an independent exponential amount of time with rate η if he has not yet been matched. The utility of a match between any buyer and any seller is a general random variable. In this centralized model, the agents make no explicit decisions, and at the time of an agent arrival, the system manager observes all matching utilities between the arrival and all potential mates (e.g., sellers if the arrival is a buyer) who are currently in the market. Using information about the number of buyers and sellers and their matching utilities, the system manager decides when to make matches and which agents to match.

Centralized dynamic matching markets occur in settings such as organ transplants, public housing,

labor markets, and various online platforms. In practice, matching utilities include information about tissue type matching and the geographical distance between the donor and the recipient for organ transplants; the location and desirability of the residence and the distance between the residence and the applicant's current residence in public housing; and the match between the needs of the employer and the experience and skills of the job applicant in the labor market. This information can lead to wide variations in the matching utilities between different buyers and sellers, and our goal is to understand how best to exploit this variation when managing the market. However, in our idealized model, the details about this information are suppressed (e.g., we do not use covariates describing the agents to help make decisions) and aggregated into the matching utility distribution between buyers and sellers.

A key issue in centralized dynamic matching markets is to find the optimal market thickness; that is, rather than match a new agent upon its arrival, it may be preferable to place the arriving agent in the market and allow more agents to arrive in the hope of making a higher-utility match in the future. In our model, we aim to maximize the long-run expected average utility rate (i.e., utility of matches per unit time) of all matches. Although we do not explicitly include agent waiting costs, a strategy that forces agents to wait too long for the market to thicken can backfire because agents may abandon the market before they are matched.

Due to the challenging nature of this problem, we resort to asymptotic methods. We consider a sequence of systems where the arrival rates in the nth system are multiplied by $n > 0$. In the absence of any matching, the number of agents of each type would be precisely the number of customers in an $M/M/\infty$ queue, which would be $O(n)$ (a generic function $f(n)$ is $O(n)$ if $\limsup_{n\to\infty} \frac{f(n)}{n} \le c$ for some finite constant $c > 0$). We use two types of asymptotic methods: one is a fluid analysis of the two-dimensional Markov process for the number of buyers and sellers in the market when the arrival rates are large. The other is extreme value theory (Gumbel 1958, Galambos 1978) and regularly varying functions (Resnick 1987), which are used because the utility of a match under the policies we consider is the maximum of a (typically) large number of random variables. In our study, a fluid analysis of the queueing process is sufficient to derive our results and leads to a decoupling of the extremal behavior of the utilities and the dynamics of the queueing system. This decoupling in turn allows us to consider correlated utilities, which is a feature that is lacking in other dynamic matching models.

In this asymptotic regime, we compute an upper bound on the utility rate of any policy that makes matches only at agent arrival epochs and compare it to the utility rate of two families of threshold policies:the population threshold policy and the utility threshold policy. Under the population threshold policy, the system manager immediately matches an arriving agent to the available mate with the highest matching utility (at which point, the arriving agent and its matched mate exit the system and their matching utility is collected by the system manager) only if the number of available mates in the market exceeds a specified threshold; otherwise, the arriving agent is not immediately matched and is instead placed in the market. Under the utility threshold policy, the arriving agent is immediately matched to its best available mate only if the corresponding matching utility exceeds a specified threshold.

Although possibly not optimal among all policies, these single-parameter policies are easy to implement and describe and allow for quite explicit results. In fact, the population threshold policy can be implemented without ever calculating the utility of individual matches (although the probability distribution of matches is required to compute the optimal threshold): all that is required is a ranked ordering of the possible matches. As we will discuss, the utility threshold policy outperforms the population threshold policy in our examples, but the latter policy is asymptotically optimal in certain cases. Another natural class of policies to consider is a batching policy, where the system manager—after a certain amount of time or after a certain number of buyers and/or sellers collect in the market—matches a set of agents. This approach requires an optimization algorithm to perform the matching and hence is more computationally demanding than our two threshold policies. Moreover, if there are many agents who abandon quickly after arrival, as in some call centers (e.g., figure 20 in Gans et al. 2003), a batching policy may not be very robust in practice. Nonetheless, in Section 7 we consider a batch-and-match policy that periodically (with an asymptotically optimal time window) optimally matches all agents on the thinner side of the market with an equal number of agents randomly selected from the thicker side of the market.

1.1. Preview of Results

In extreme value theory, the limiting distribution of the maximum of many random variables can be one of three types, loosely based on whether the underlying distribution of these random variables has an exponential right tail, has a heavier (e.g., power law) right tail, or is bounded from above, and our results are qualitatively different in each case. Although our main results are couched in terms of regularly varying functions, we preview our results with three canonical examples (Table 1)—one from each of the three

domains of attraction—in the symmetric case, which are analyzed in Section 10 of the online appendix. When matching utilities have an exponential distribution, the population threshold policy with a threshold of $\frac{n}{\ln n}$ is asymptotically optimal with a utility rate that is $O(n \ln n)$ and twice as large as the utility rate of the greedy policy—that is, the population threshold policy with a threshold of zero—in the limit. When the matching utilities have a Pareto (c, β) distribution with shape parameter $\beta > 1$ (and hence a finite mean), the population threshold $\frac{\lambda}{\eta(1+\beta)} n$ is asymptotically optimal. Although the utility rate of this threshold policy does not converge to the loose upper bound in this case, the utility rate and the upper bound are both $O(n^{1+1/\beta})$, whereas the utility rate under the greedy policy is only $O(n^{1+1/(2\beta)})$. When the matching utilities have a uniform distribution, the greedy policy is asymptotically optimal (i.e., 0 is an asymptotically optimal population threshold) and the optimal utility rate is $O(n)$.

In the Pareto case, the utility threshold $0.763\sqrt{\pi n}$ is asymptotically optimal when $c = 1$ and $\beta = 2$, and the corresponding utility rate is $O(n^{1+1/\beta})$. This asymptotic utility rate is computed explicitly, and it is shown to be larger than the utility rate of the asymptotically optimal population threshold policy. In the exponential and uniform cases, where we have already identified an asymptotically optimal policy, we use heuristics to compute, in the prelimit, utility threshold policies that are consistent with the asymptotically optimal descriptions but outperform in simulation results the population threshold policy. We also consider a positively correlated Pareto case in Section 10.4 of the online appendix, and show that the asymptotically optimal population threshold is independent of the correlation, the asymptotically optimal utility threshold decreases as the correlation increases, and the utility rates of both threshold policies decrease as the correlation increases. In Section 6, we consider an unbalanced market, where buyers have a different arrival rate and abandonment rate than sellers, and we analyze the utility threshold policy in the heavy-tailed case. Surprisingly, although we allow the buyers and sellers to have a different utility threshold, we find that they have the same asymptotically optimal utility threshold. Finally, we show in Section 7 that in the Pareto case, the utility threshold policy outperforms the batch-and-match policy.

Taken together, the optimal amount of patience—and hence market thickness—increases with the right tail of the matching utility distribution, as does the optimal utility rate and the performance gap between the utility threshold policy and the population threshold policy. Our limited analysis of an unbalanced market suggests that the optimal market thickness also increases with the amount of imbalance. In our particular model of correlation, increased positive correlation among matching utilities decreases the benefit of increased patience (i.e., the system manager is less likely to observe a future utility that is much better than the best existing utility), whereas the cost of increased patience (i.e., the number of abandonments) is independent of the correlation. Among the three one-parameter policies considered here, the utility threshold policy displays the best performance.

1.2. Related Work

Matching markets is a large and active area of research, and we restrict our review to centralized dynamic markets. Although our model lacks the contextual richness of some of the models for specific types of markets, the most distinctive feature of our model is the generality of the matching utilities, which allows us to understand how the right tail of the matching utility distribution impacts the optimal thickness of the market (Table 1). In contrast, much of the recent work in dynamic (centralized or decentralized) matching markets, either via two-type agents (e.g., easy-to-match or hard-to-match agents, or matches that are preferred or nonpreferred; Baccara et al. 2020, Ashlagi

Table 1. Summary of Results for the Three Canonical Cases in Section 10 of the Online Appendix

	Matching utility distribution		
Policy	Exponential(ν)	Pareto(c, shape $\beta > 1$)	Uniform(a, b)
Upper bound	$U_n^+ \sim \frac{\lambda}{\nu} n \ln n$	$U_n^+ = O(n^{1+1/\beta})$	$U_n^+ \sim \lambda b n$
Greedy policy	$U_n^g \sim \frac{\lambda}{2\nu} n \ln n$	$U_n^g = O(n^{1+1/(2\beta)})$	Asymptotically optimal
Population threshold policy with threshold z_n	$z_n^* = \frac{n}{\ln n}$ is asymptotically optimal	$z_n^* = \frac{\lambda}{\eta(1+\beta)} n$ $U_n^p(z_n^*) = O(n^{1+1/\beta})$ but no convergence to upper bound unless $\beta \to \infty$	$z_n^* = 0$ is asymptotically optimal
Utility threshold policy with threshold v_n	Heuristic $v_n^* = \frac{\ln n - \ln\ln n - \ln\ln(\frac{2\lambda \ln n}{\lambda \ln n + \eta})}{\nu}$	$v_n^* = 1.353\sqrt{n}$ when $c = 1$, $\beta = 2$; $U_n^u(v_n^*) = O(n^{1+1/\beta})$	heuristic $v_n^* = a +$ $(b-a)\left(\frac{1}{\sqrt{2n\pi}} + \frac{1}{2}\right)^{\frac{\eta}{\lambda}\sqrt{\frac{\pi}{2n}}}$

Note. U_n^+, U_n^g, $U_n^p(z_n)$, and $U_n^u(v_n)$ are, respectively, the upper bound on the utility rate for any arrival-only policy, the utility rate for the greedy policy, the utility rate for the population threshold policy with threshold z_n, and the utility rate for the utility threshold policy with threshold v_n, all for the nth system.

et al. 2018a, 2019) or a compatibility network (Ashlagi et al. 2013, Anderson et al. 2017, Varma et al. 2019, Akbarpour et al. 2020) essentially lead to dichotomous outcomes for a match. Exceptions include Ünver (2010), who considers blood type compatibility for a dynamic kidney exchange model, Emek et al. (2016) and Ashlagi et al. (2017b), who consider minimizing mismatch costs when agents arrive on a finite metric space in a nonbipartite and bipartite setting, respectively, and Ashlagi et al. (2018b), who allow general matching utilities in a discrete time model with a constant time until abandonment. They perform a primal-dual analysis to derive competitive ratios for algorithms when there is no prior information about the match values or arrival times.

The analysis of multiclass matching queues is an active area. Hu and Zhou (2021) consider a discrete-time, multiclass, discounted variant of our problem that includes waiting costs. They show that the optimal policy is of threshold form under vertical and unidirectionally horizontal differentiated types. Ding et al. (2021) allow the matching utilities to depend on the class of buyer and seller and perform a fluid analysis of a greedy policy, and Bušić and Meyn (2015) minimize linear holding costs in a system without class-dependent matching utilities or abandonment but also find that matches are not made until there is a sufficient number of agents in the market (see Moyal and Perry 2017, where these systems are referred to as *matching queues*, for other references to these types of models).

Gurvich and Ward (2014) and Nazari and Stolyar (2019) study a control problem in a more general setting than the aforementioned studies, where arriving customers wait to be matched to agents of other classes. Gurvich and Ward (2014) minimize cumulative holding costs over a finite horizon and show that a myopic discrete-review matching algorithm is asymptotically optimal. Nazari and Stolyar (2019) maximize the long-run average revenue rate subject to maintaining stable queues and construct a greedy primal-dual approach that is asymptotically optimal. It is difficult to compare this powerful result to our results, given that we assume abandonment rather than stability, and we have a single-class model with general rewards rather than a multiclass model with class-dependent rewards.

Two other studies consider fluid and diffusion limits of simplified versions of our model where either a match occurs with a certain probability for each buyer-seller pair (Büke and Chen 2017) or everyone matches when there is an available mate (Liu et al. 2015), which corresponds to our greedy policy but with a deterministic utility (i.e., a matching utility distribution that is a point mass at one value). In both cases, the system state reduces to a one-dimensional quantity (the number of sellers minus the number of buyers), whereas our model requires a two-dimensional state space for a nongreedy policy.

Perhaps the most closely related paper is Mertikopoulos et al. (2020), which also considers a symmetric centralized dynamic matching market. Compared with our study, they assume independent exponential mismatch costs rather than general matching utilities, consider waiting times rather than abandonment, and are interested in minimizing the sum of mismatch and waiting costs over a finite horizon. They consider a class of policies that make the kth match (which has the lowest mismatch cost among possible matches) when the short side of the market grows to a certain one-parameter function of k. They analyze the performance of the policy (using the celebrated $\pi^2/6$ result for the expected minimum weight matching due to Mezard and Parisi 1987 and rigorously proved by Aldous 2001) under various values of the parameter and also identify a policy that balances the mismatch and waiting costs. It is difficult to draw qualitative comparisons between our results for exponential utilities (which incorporate abandonments) and their results (which incorporate waiting costs); indeed, our approach depends on the right tail of the exponential distribution via extreme value theory, whereas their approach depends on the left tail of the exponential distribution via minimum weighted matching.

We briefly mention other works that are only peripherally related. Originally motivated by public housing (Kaplan 1988), Caldentey et al. (2009) and Adan and Weiss (2012) consider infinite bipartite matching of servers and customers under the first-come first-served policy. There is also a stream of work in online bipartite matching in an adversarial setting (Karp et al. 1990), where agents do not wait in the market if they are not matched immediately. Finally, there is a body of literature (e.g., Duffie et al. 2018 and references therein) that uses the law of large numbers to analyze the performance of static and dynamic matching models used in economics, finance, and genetics, but these models are descriptive rather than prescriptive.

1.3. Organization

The paper is organized as follows. We formulate the model in Section 2 and state our main theoretical results in Section 3, which are proved in Section 9 of the online appendix. After analyzing a greedy policy in Section 4, we apply our main results to specific matching utility distributions in Section 10 of the online appendix and assess the accuracy of these results in a simulation study in Section 5. The unbalanced case is studied in Section 6, the batch-and-match policy is analyzed in Section 7, and concluding remarks are offered in Section 8.

1.4. Notation

For the convenience of the reader, we collect together the notational conventions used in this paper. Although we have already introduced the notation $O(n)$, we repeat it here: a generic function $f(n)$ is $O(n)$ if $\limsup_{n\to\infty}\frac{f(n)}{n}\le c$ for some finite constant $c>0$. In a similar vein, we introduce $o(n)$, $\Omega(n)$, and $\Theta(n)$. A generic function $f(n)$ is $o(n)$ if $\lim_{n\to\infty}\frac{f(n)}{n}=0$, is $\Omega(n)$ if there exist $c>0$ and an integer $n_o\ge 1$ such that $f(n)\ge cn$ for all integers $n\ge n_o$, and is $\Theta(n)$ if $f(n)$ is both $O(n)$ and $\Omega(n)$. We use $x_n\sim y_n$ as shorthand for $\frac{x_n}{y_n}\to 1$ as $n\to\infty$.

We let $\mathbb{R}$ denote the real line, and, for any finite integer $k\ge 1$, we let $\mathbb{R}^k$ denote the k-dimensional Euclidean space. The Euclidean norm of $x\in\mathbb{R}^k$ is denoted by $|x|$. We let $\mathbb{R}_+$ denote the set of nonnegative reals, and we let $\mathbb{Z}_+$ denote the set of nonnegative integers. The stochastic processes that we consider take values in $\mathbb{R}^k$ and are assumed to be elements of $\mathbb{D}^k[0,\infty)$, the space of right continuous functions mapping $[0,\infty)$ into $\mathbb{R}^k$ that have left limits, endowed with the Skorokhod topology.

For $x\in\mathbb{R}_+$, $\lceil x\rceil$ is the smallest integer that is not smaller than x. The standard stochastic order between two distribution functions F_1 and F_2 is denoted by $F_1\le_{st}F_2$. We use $\stackrel{d}{=}$ to denote equality in distribution. More specifically, we write $X\stackrel{d}{=}\text{Poisson}(x)$ to denote that the random variable X has a Poisson distribution with mean x.

2. The Model

2.1. Dynamics

Buyers and sellers arrive to the market according to independent Poisson processes with rate λ. The agents are impatient, in that each buyer and each seller independently abandons the market after an independent and identically distributed (i.i.d.) exponential amount of time with rate η if they are not matched within this time. If an agent is matched prior to his abandonment, then the agent leaves at the time of matching.

Let $B(t)$ and $S(t)$ be the number of buyers and sellers in the system at time t; these agents have arrived but have not yet abandoned or been matched. The utility of a match between any buyer and any seller is a random variable $V\ge 0$ with cumulative distribution function (CDF) $F(v)$. When a buyer (seller, respectively) arrives to this centralized system to find it in state $(B(t),S(t))$, then $S(t)$ ($B(t)$, respectively) instances of V are observed by the system manager, which represent the matching utilities of the arriving agent with all currently available potential mates. Thus, at any point in time, the system manager knows the utility that would be generated by matching any buyer to any seller.

2.2. Policies

Our goal is to maximize the long-run expected average rate of utility from matches, which we refer to as the *utility rate*. Whereas the system manager could conceivably make matches at any point in time, we restrict our attention to *arrival-only policies*, where a match may occur only at the arrival epoch of one of the agents being matched. In particular, we consider the following two classes of arrival-only policies.

1. Population threshold policies: A buyer who arrives at time t is matched immediately to a seller if the number of sellers in the system satisfies $S(t)\ge z$; in this case, the arriving buyer is matched to the seller who has the highest matching utility with the buyer, with ties broken arbitrarily. If $S(t)<z$, then the arriving buyer waits in the market and leaves upon being matched to a later-arriving seller or upon abandonment. Similarly, a seller who arrives at time t is immediately matched to the highest-matching buyer if $B(t)\ge z$ and waits otherwise. We refer to the parameter z as the *population threshold*.

2. Utility threshold policies: A buyer who arrives at time t is matched immediately to the seller with matching value $\max_{1\le i\le S(t)}V_i$ if $\max_{1\le i\le S(t)}V_i>v$ for some fixed $v\ge 0$, with ties broken arbitrarily. If $\max_{1\le i\le S(t)}V_i\le v$, then the arriving buyer waits in the market and leaves upon being matched to a later-arriving seller or upon abandonment. Similarly, a seller who arrives at time t is immediately matched to the buyer with matching value $\max_{1\le i\le B(t)}V_i$ if $\max_{1\le i\le B(t)}V_i>v$ and waits otherwise. We refer to the parameter v as the *utility threshold*.

Given the symmetry of the underlying stochastic model, it seems natural to restrict ourselves to single-parameter policies, where buyers and sellers have the same threshold (z or v). In the analysis of the unbalanced case in Section 7, we allow different utility thresholds for buyers and sellers (v_b and v_s) and find that the asymptotically optimal values satisfy $v_b=v_s$ under Pareto matching utilities. This result suggests that a single-parameter threshold policy is not only easier to use in practice and easier to analyze than a two-parameter threshold policy, but it also does not sacrifice performance.

2.3. Utilities

In our model, the utilities of potential matches of a new arrival with agents on the other side of the market may be correlated. However, we make the following assumption.

Assumption 1. *There exists a sequence of distributions $F_1, F_2, \ldots$ such that $F_{k-1} \leq_{st} F_k$ for each k and, for an arriving agent who finds k agents on the other side of the market, $\max\{V_1, \ldots, V_k\}$ is independent of the past and has distribution F_k.*

For example, if the utilities of different matches are i.i.d. with distribution F, then $F_k(x) = (F(x))^k$ in Assumption 1. But Assumption 1 allows us to deal with correlated utilities, which is natural when there is contextual information (e.g., covariates) that can be used to inform the utilities based on the types of buyers and sellers to be matched. More specifically, Assumption 1 holds if the utilities are conditionally independent given a context observed at the time of arrival. In this case, the equation $F_k(x) = (F(x))^k$ holds with an additional expectation, and the stochastic ordering in k still holds.

Let the random variable $M(k) \triangleq \max\{V_1, \ldots, V_k\}$ have distribution F_k. We impose the following assumption on $M(k)$.

Assumption 2. *For each $x \in \mathbb{R}_+$, define*

$$m(x) = E[M(\lceil x \rceil)],$$

and suppose that $m(\cdot)$ is regularly varying with index $\alpha \in [0,1)$. That is, for every $x > 0$,

$$\lim_{t\to\infty} \frac{m(tx)}{m(t)} = x^{\alpha}. \tag{1}$$

A regularly varying function with index $\alpha = 0$ is also known as slowly varying.

For the case of i.i.d. utilities, Assumption 2 covers every utility distribution such that $E(V^{1+\delta}) < \infty$ for some $\delta > 0$. All distributions that belong to the maximum domain of attraction of a generalized extreme value distribution—which unifies the type I (Gumbel), type II (Frechet), and type III (Weibull) laws within a single parametric family—satisfy (1) (including, e.g., uniform, beta, gamma, lognormal, and Pareto). There are also other distributions that do not belong to any domain of attraction in extreme value theory for which (1) holds; for example, the geometric, negative binomial, and Poisson distributions satisfy (1) with $\alpha = 0$. For ease of reference, we collect some basic facts about extreme value theory and regularly varying functions in Section 11 of the online appendix.

The case $\alpha = 0$ corresponds to distributions for which all moments exist (i.e., the tail of V decays faster than any polynomial), whereas $\alpha > 0$ corresponds to the case in which the tails of V decrease roughly like a polynomial with degree $1/\alpha$. The condition that $\alpha < 1$ is imposed to guarantee that $E(V^{1+\delta}) < \infty$ for some $\delta > 0$. We will refer to $\alpha = 0$ as the *light-tailed case* and $\alpha \in (0,1)$ as the *heavy-tailed case*.

2.4. Scaling

To make further progress, we consider a sequence of systems indexed by $n = 1, 2, \ldots$, and some quantities in the nth system include the subscript n. The arrival rate in the nth system is $n\lambda$, and the abandonment rate in the nth system is η. Alternatively and equivalently, we could leave the arrival rate unscaled and slow down the abandonment rate by a factor of n, as in Liu et al. (2015). The matching utilities are unscaled. In the nth system, we denote the system state by $(B_n(t), S_n(t))$, the population threshold by z_n, the utility threshold by v_n, and the utility rate by U_n.

3. Main Results

Results for the population threshold policy and the utility threshold policy are given in Theorem 1 in Section 3.1 and in Theorem 2 in Section 3.2, respectively. Theorem 1 shows that the optimal population threshold policy is asymptotically optimal among the class of arrival-only policies when $\alpha = 0$ and provides the asymptotically optimal population threshold when $\alpha \in (0,1)$. Theorem 2 provides the asymptotically optimal utility threshold when $\alpha \in (0,1)$. The proofs of Theorems 1 and 2 appear in Section 9 of the online appendix.

3.1. Population Threshold Policy

We begin by providing a dynamic description of the system using Poisson processes. Denote the indicator function of event x by $I_{\{x\}}$, and let $N_B^+(\cdot), N_B^-(\cdot), N_S^+(\cdot), N_S^-(\cdot)$ be independent Poisson processes with unit rate, which are used to construct buyer arrivals, buyer abandonments, seller arrivals, and seller abandonments, respectively. Under the population threshold policy with threshold z_n, the state (B_n, S_n) of the nth system at time t satisfies

$$B_n(t) = B_n(0) + \int_0^t I_{\{S_n(r_-) < z_n\}} dN_B^+(\lambda n r) - N_B^-\left(\eta \int_0^t B_n(r)dr\right) - \int_0^t I_{\{B_n(r_-) \geq z_n\}} dN_S^+(\lambda n r), \tag{2}$$

$$S_n(t) = S_n(0) + \int_0^t I_{\{B_n(r_-) < z_n\}} dN_S^+(\lambda n r) - N_S^-\left(\eta \int_0^t S_n(r)dr\right) - \int_0^t I_{\{S_n(r_-) \geq z_n\}} dN_B^+(\lambda n r). \tag{3}$$

The process $\{B_n(t), S_n(t), t \geq 0\}$ is a nonnegative (entry wise) irreducible two-dimensional birth-and-death process on a subset of $\mathbb{Z}_+ \times \mathbb{Z}_+$, and each coordinate is bounded by that of an infinite-server queue, for each $n > 0$. Thus, the process $\{B_n(\cdot), S_n(\cdot)\}$ is a positive-recurrent continuous-time Markov chain and therefore it possesses a stationary distribution, which we

denote by $(B_n(\infty), S_n(\infty))$. By symmetry, the utility rate $U_n^p(z_n)$ of the population threshold policy with threshold z_n can be expressed as

$$\begin{aligned} U_n^p(z_n) &= \lambda n E\left[m(B_n(\infty))I_{\{B_n(\infty)\geq z_n\}}\right] \\ &\quad + \lambda n E\left[m(S_n(\infty))I_{\{S_n(\infty)\geq z_n\}}\right], \\ &= 2\lambda n E\left[m(B_n(\infty))I_{\{B_n(\infty)\geq z_n\}}\right]. \end{aligned} \tag{4}$$

The next theorem shows that, for $\alpha = 0$, the population threshold policy is asymptotically optimal among the family of arrival-only policies. Also, for each $\alpha \in [0,1)$, it characterizes the scaling of the optimal population threshold, that is, the threshold z_n that maximizes the utility rate asymptotically as $n \to \infty$.

Theorem 1. *Suppose that Assumption 2 holds.*

(i) *If $\alpha = 0$, then there exists an $o(n)$ sequence of population thresholds z_n^* such that $\lim_{n\to\infty} \frac{m(n)}{m(z_n^*)} = 1$. For any such sequence of thresholds, the population threshold policy is asymptotically optimal in the following sense. Let $U_n^p(z_n^*)$ and U_n be the utility rates under the aforementioned policy and any other arrival-only policy, respectively. Then $\liminf_{n\to\infty} \frac{U_n^p(z_n^*)}{U_n} \geq 1$. The associated utility rate satisfies*

$$\lim_{n\to\infty} \frac{U_n^p(z_n^*)}{nm(n)} = \lambda. \tag{5}$$

(ii) *If $\alpha \in (0,1)$, then the population threshold policy with $z_n^* = z_* n$, where $z_* = \frac{\lambda\alpha}{\eta(1+\alpha)}$ is asymptotically optimal among the class of population threshold policies. The associated utility rate satisfies*

$$\lim_{n\to\infty} \frac{U_n^p(z_n^*)}{nm(n)} = \lambda z_*^\alpha \left(1 - \frac{\eta z_*}{\lambda}\right). \tag{6}$$

For $\alpha = 0$, it remains to compute an $o(n)$ sequence of thresholds z_n^* such that $\lim_{n\to\infty} \frac{m(n)}{m(z_n^*)} = 1$. This is usually not difficult to do. For example, when utilities are i.i.d. with an exponential distribution, then $z_n^* = \frac{n}{\ln n}$ satisfies this property. More generally, as shown in theorem 1 in Bojanic and Seneta (1971), for a large class of distributions, setting $z_n^* = \frac{n}{m(n)^\delta}$ for any positive real δ is sufficient. By setting z_n in this way (i.e., $o(n)$ but not too small), we simultaneously ensure the following: (1) the fraction of agents that abandon the system tends to 0, and (2) the market thickness, that is, $B_n(\infty)$, is almost linear in n. In other words, almost all agents experience maximal utility. This can be seen most clearly in Equation (5), where the utility rate under the optimal population threshold policy satisfies $U_n^p(z_n^*) \sim n\lambda m(n)$, which is the arrival rate of buyers times the expected value of the maximum of n matching utilities.

However, for heavy-tailed distributions in part (ii) of Theorem 1, $m(z_n)$ for any $o(n)$ sequence z_n is vanishingly small compared with $m(n)$. Thus, it is not possible to ensure that most users see maximal utility, implying that our simple upper bound is unachievable. Moreover, to maximize the utility rate, it is not obvious whether the system manager should set $z_n = o(n)$ to guarantee that most agents are matched instantly or should set $z_n = O(n)$ to ensure that market thickness is maximal even if a nontrivial fraction of users abandon the system. Part (ii) of Theorem 1 implies that the latter option is the right choice under heavy-tailed distributions.

We conclude this subsection with a brief sketch of the proof of Theorem 1, which relies on a fluid analysis of Equations (2)–(3). We define $\bar{B}_n(t) = n^{-1}B_n(t)$ and $\bar{S}_n(t) = n^{-1}S_n(t)$. Because the formal limit of $(B_n(t), S_n(t))$ involves indicator functions that are not continuous (see (35)–(36) in Section 9.2 of the online appendix), we need to study the limiting dynamical system as the solution to the following Skorokhod problem:

$$\bar{B}(t) = \bar{B}(0) + \lambda t - \eta \int_0^t \bar{B}(r)dr - L_z^{\bar{B}}(t) - L_z^{\bar{S}}(t), \tag{7}$$

$$\bar{S}(t) = \bar{S}(0) + \lambda t - \eta \int_0^t \bar{S}(r)dr - L_z^{\bar{B}}(t) - L_z^{\bar{S}}(t), \tag{8}$$

where $L_z^{\bar{B}}(\cdot)$, $L_z^{\bar{S}}(\cdot)$ are nondecreasing processes such that $L_z^{\bar{B}}(0) = L_z^{\bar{S}}(0) = 0$ and

$$\int_0^t (\bar{B}(r) - z)dL_z^{\bar{B}}(r) = \int_0^t (\bar{S}(r) - z)dL_z^{\bar{S}}(r) = 0, \tag{9}$$

and $\bar{B}(t), \bar{S}(t) \leq z$. To obtain explicit expressions for the Skorokhod problem, we use the change of variables $\bar{B}_z(t) = z - \bar{B}(t)$, $\bar{S}_z(t) = z - \bar{S}(t)$, and $\bar{\lambda}_z = \lambda/\eta - z \geq 0$. This allows us to reduce (9) to the one-dimensional condition

$$\int_0^t \min(\bar{B}_z(r), \bar{S}_z(r))dL(r) = 0, \quad \mathrm{L}(0) = 0,$$

which enables us to obtain an explicit solution to (7)–(9). With this solution in hand, we show uniqueness and then apply a standard Picard iteration to argue existence.

We use martingale arguments to show that $(\bar{S}_n(\cdot), \bar{B}_n(\cdot)) \to (\bar{S}(\cdot), \bar{B}(\cdot))$ uniformly on compact sets in probability. The dynamical system describing $(\bar{B}, \bar{S})$ has the unique attractor (z, z) if $\lambda/\eta \geq z$, given the initial condition $\bar{B}(0) \leq z$, $\bar{S}(0) \leq z$. We then show that the limit interchange ($t \to \infty$ and $n \to \infty$) holds, and we prove that $(\bar{B}_n(\infty), \bar{S}_n(\infty)) \to (z, z)$ almost surely as $n \to \infty$.

The next step in the proof is to compute the utility rate. Taking expectations on both sides of Equation (2) yields

$$\eta E\left[\bar{B}_n(\infty)\right] = \lambda\{P(\bar{S}_n(\infty) < z) - P(\bar{B}_n(\infty) \geq z)\}, \tag{10}$$

from which we can obtain, using symmetry arguments, that

$$\lim_{n\to\infty} P(\bar{B}_n(\infty) \geq z) = \frac{1}{2}\left(1 - \frac{\eta z}{\lambda}\right). \tag{11}$$

The following key lemma, which is proved in Section 9 of the online appendix, allows us to compute the utility rate in (4). Recall that $m(n)$ is defined in Assumption 2.

Lemma 1. *Let $\{N_n\}_{n\geq 1}$ be a sequence of positive random variables taking values on the positive integers, and let $\bar{N}_n = E(N_n) < \infty$. Assume that $\bar{N}_n \to \infty$ and that $P(|N_n - \bar{N}_n| > \varepsilon \bar{N}_n) \to 0$. Then $E[m(N_n)] \sim m(\bar{N}_n)$ as $n \to \infty$.*

Using Lemma 1 and Equation (11) and setting $z_n = nz$ allows us to compute the utility rate

$$U_n^p(z_n) = \lambda n m(zn)\left(1 - \frac{\eta z}{\lambda}\right)(1 + o(1)) \tag{12}$$

as $n \to \infty$, and combining (12) with Equation (1) yields

$$\frac{U_n^p(z_n)}{nm(n)} = \lambda z^{\alpha}\left(1 - \frac{\eta z}{\lambda}\right)(1 + o(1)). \tag{13}$$

In the $\alpha \in (0,1)$ case, we optimize the right-hand side of (13) with respect to z to obtain the asymptotically optimal population threshold $z_n^* = z_* n$, where $z_* = \frac{\lambda\alpha}{\eta(1+\alpha)}$.

In the $\alpha = 0$ case, we similarly use the fluid limit analysis to show that $E[B_n(\infty)]$ is $o(n)$ for any sequence of thresholds z_n that is $o(n)$. Furthermore, the arguments used to obtain (11) also imply that

$$\lim_{n\to\infty} P(B_n(\infty) \geq z_n) = \lim_{n\to\infty} P(S_n(\infty) \geq z_n) = \frac{1}{2}. \tag{14}$$

A PASTA (Poisson arrivals see time averages) argument implies that $U_n^p(z_n) \geq \lambda n m(z_n)(1 + o(1))$. Consequently, for any sequence $z_n = o(n)$ such that

$$\lim_{n\to\infty} \frac{m(z_n)}{m(n)} = 1,$$

we would have that $U_n^p(z_n) \geq \lambda n m(n)(1 + o(1))$. Lemma 3 in Section 9.1 of the Appendix guarantees that such a sequence exists.

Finally, asymptotic optimality in part (i) of Theorem 1 follows from the aforementioned results by constructing the following simple upper bound (see Section 9 of the online appendix for a proof of Lemma 2) on the performance of any arrival-only policy, which uses Lemma 1 and assumes that all agents are matched (and hence the arrival rate in Lemma 2 is λn) and that—when computing $B_n(\infty)$ in Equation (4)—agents leave only upon abandonment (implying that $B_n(\infty) \overset{d}{=} \text{Poisson}(\lambda n/\eta)$).

Lemma 2. *Let U_n be the utility rate for any arrival-only policy. Then an upper bound U_n^+ is given by*

$$U_n \leq U_n^+ = \lambda n m\left(\frac{\lambda n}{\eta}\right).$$

3.2. Utility Threshold Policy

Because the population threshold policy is asymptotically optimal within the class of arrival-only policies when $\alpha = 0$, we focus on the case $\alpha \in (0,1)$ in Theorem 2. In order to describe the dynamics of the utility threshold policy, we introduce two independent arrays of nonnegative i.i.d. random variables, $\{V_{i,j}^B : i \geq 1, j \geq 1\}$ and $\{V_{i,j}^S : i \geq 1, j \geq 1\}$, having CDF $F(\cdot)$. We let $\{A_j^B : j \geq 1\}$ be the sequence of arrival times associated with the process $N_B^+(n\lambda\cdot)$, and we let $\{A_j^S : j \geq 1\}$ be the sequence of arrival times associated with the process $N_S^+(n\lambda\cdot)$. The dynamics can be described path-by-path as follows:

$$\begin{aligned} B_n(t) &= B_n(0) + \sum_{j=1}^{N_B^+(n\lambda t)} I_{\{\max_{i=1}^{S_n(A_{j-}^B)} V_{i,j}^B \leq v\}} - \sum_{j=1}^{N_S^+(n\lambda t)} I_{\{\max_{i=1}^{B_n(A_{j-}^S)} V_{i,j}^S > v\}} \\ &\quad - N_B^-\left(\eta \int_0^t B_n(r_-)dr\right), \\ S_n(t) &= S_n(0) + \sum_{j=1}^{N_S^+(n\lambda t)} I_{\{\max_{i=1}^{B_n(A_{j-}^S)} V_{i,j}^S \leq v\}} - \sum_{j=1}^{N_B^+(n\lambda t)} I_{\{\max_{i=1}^{S_n(A_{j-}^B)} V_{i,j}^B > v\}} \\ &\quad - N_S^-\left(\eta \int_0^t S_n(r)dr\right). \end{aligned} \tag{15}$$

By symmetry and ergodicity, we can express the utility rate $U_n^u(v_n)$ for the utility threshold policy with threshold v_n as

$$U_n^u(v_n) = 2\lambda n E[E[M(B_n(\infty)) I_{\{M(B_n(\infty)) \geq v_n\}} | B_n(\infty)]]. \tag{16}$$

Because the analysis of the utility threshold policy considers the entire distribution of the maximum rather than only its expected value, we need to strengthen Assumption 2 by imposing the following additional assumption.

Assumption 3. *In addition to Assumption 2, suppose that $\alpha \in (0,1)$ and*

$$\frac{M(n)}{m(n)} \Rightarrow X \quad \text{as } n \to \infty,$$

where $P(X > t) = 1 - e^{-\kappa/t^{1/\alpha}}$ and κ is a normalizing constant such that $E(X) = 1$.

That is, $X = (\kappa^{-1}T)^{-\alpha}$ is an exponential random variable with mean one. Assumption 3 is satisfied if the

utilities belong to the domain of attraction of the Frechet law, which in turn is equivalent, in the i.i.d. case, to requiring the distribution of utilities to be regularly varying with index $1/\alpha$ (see section 1.2, proposition 1.11 of Resnick 1987).

Theorem 2. *Suppose that Assumption 3 holds. For* $x \in [0, \lambda/\eta]$, *define*

$$v(x) = \left(\frac{\kappa x}{\ln\left(\frac{2\lambda}{\eta x + \lambda}\right)} \right)^{\alpha}.$$

Then there exists a unique solution $x_* \in (0, \lambda/\eta)$ *satisfying*

$$x_*^{1-\alpha} v(x_*) \frac{\eta}{2\lambda\alpha\kappa^{\alpha}} = \int_0^{\kappa x_*/v(x_*)^{1/\alpha}} t^{-\alpha} e^{-t} dt.$$

Moreover, a threshold policy with utility threshold $v_n^* = v(x_*)m(n)$ *is asymptotically optimal among the class of utility threshold policies and the associated utility rate satisfies*

$$\lim_{n\to\infty} \frac{U_n^u(v_n^*)}{nm(n)} = 2\lambda x_*^{\alpha} E\left[XI_{\left\{X \geq \frac{v(x_*)}{x_*^{\alpha}}\right\}} \right].$$

As in the population threshold policy, this result shows that, for heavy-tailed distributions, it is beneficial to ensure that market thickness is maximal at the cost of abandonment of a nontrivial fraction of users in the system. Although we do not prove any results for the utility threshold policy in the $\alpha = 0$ case (since asymptotic optimality is already achieved for the population threshold policy), we show in Section 10 of the online appendix how heuristics inspired by Theorems 1 and 2 can lead to effective utility thresholds in the $\alpha = 0$ case.

The proof of Theorem 2 uses the same general approach as in the proof of part (ii) of Theorem 1, and we briefly outline it here. We assume that the thresholds satisfy

$$\frac{v_n}{m(n)} \to v \text{ for some } v \geq 0,$$

and we use Assumption 3 to show that the putative fluid limit of $B_n(t) = n^{-1}B_n(t)$ and $S_n(t) = n^{-1}S_n(t)$ is

$$\bar{B}(t) = \bar{B}(0) + \lambda \int_0^t e^{-\kappa \bar{S}(r)/v^{1/\alpha}} - \eta \int_0^t \bar{B}(r) dr - \lambda \int_0^t \left(1 - e^{-\kappa \bar{B}(r)/v^{1/\alpha}}\right) dr, \tag{17}$$

$$\bar{S}(t) = \bar{S}(0) + \lambda \int_0^t e^{-\kappa \bar{B}(r)/v^{1/\alpha}} - \eta \int_0^t \bar{S}(r) dr - \lambda \int_0^t \left(1 - e^{-\kappa \bar{S}(r)/v^{1/\alpha}}\right) dr. \tag{18}$$

A martingale decomposition similar to that given in the proof of part (ii) of Theorem 1 shows that $\bar{B}_n(\cdot) \to \bar{B}(\cdot)$ and $\bar{S}_n(\cdot) \to \bar{S}(\cdot)$ uniformly on compact sets in probability. Because (17)–(18) do not pose the degeneracies involving the Skorokhod map encountered in the case of part (ii) of Theorem 1, we can use theorem 7.2 of chapter 3 in Ethier and Kurtz (2005) to show that the family $\left\{\left((\bar{B}_n(t) : t \geq 0), (\bar{S}_n(t) : t \geq 0)\right)\right\}_{n\geq 1}$ is tight in the Skorokhod topology.

The unique solution to the fluid limit satisfies

$$0 = -\eta\bar{x} - \lambda + 2\lambda e^{-\kappa\bar{x}/v^{1/\alpha}},$$

which can be expressed as

$$v(\bar{x}) = \left(\frac{\kappa\bar{x}}{\ln\left(\frac{2\lambda}{\eta\bar{x}+\lambda}\right)} \right)^{\alpha} \tag{19}$$

or

$$\bar{x}(v) = -\frac{\lambda}{\eta} + \frac{v^{1/\alpha}}{\kappa} W\left(\frac{2\lambda\kappa}{\eta v^{1/\alpha}} \exp\left(\frac{\lambda\kappa}{\eta v^{1/\alpha}} \right) \right),$$

where $W(x)$ is the Lambert W function.

We use Assumptions 2 and 3 to optimize the utility rate with respect to $\bar{x}(v)$, yielding the optimization problem

$$\sup_{\bar{x}\in(0,\lambda/\eta)} 2\lambda\bar{x}^{\alpha}\kappa^{\alpha} \int_0^{\kappa\bar{x}/v(\bar{x})^{1/\alpha}} t^{-\alpha} e^{-t} dt. \tag{20}$$

The solution to (20) reduces to $\bar{x}^*$ uniquely satisfying

$$\frac{v(\bar{x})\eta}{2\lambda\kappa^{\alpha}} = \alpha\bar{x}^{\alpha-1} \int_0^{\kappa\bar{x}/v(\bar{x})^{1/\alpha}} t^{-\alpha} e^{-t} dt,$$

and substituting $\bar{x}^*$ into (19) gives the optimal utility threshold.

4. A Greedy Policy

In Section 10 in the online appendix, we apply the results in Theorems 1 and 2 to several different matching utility distributions and then assess the accuracy of these analyses via simulation in Section 5. To provide a natural benchmark for comparison, we first analyze the greedy policy, which corresponds to the population threshold policy with threshold $z_n = 0$. That is, under the greedy policy, each arriving agent is matched to the available mate with the highest matching utility and waits in the market if there are no available mates.

Under the greedy policy, the state of the nth system can be described by $B_n(t) - S_n(t)$ because there are never both buyers and sellers in the system at the same time. By theorem 4.5 in Liu et al. (2015), the

steady-state distribution of $\frac{B_n(t)-S_n(t)}{\sqrt{n}}$ converges to $N(0, \lambda/\eta)$ as $n \to \infty$.

The probability that a buyer or seller abandons is the long-run expected number of abandonments per unit time divided by the total arrival rate of agents (i.e., buyers plus sellers), which can be approximated by

$$\frac{\sqrt{n}\eta E[|N(0,\lambda/\eta)|]}{2\lambda n} = \frac{\sqrt{n}\eta\sqrt{\frac{2}{\pi}\frac{\lambda}{\eta}}}{2\lambda n}, \tag{21}$$

$$= \frac{1}{\sqrt{2\pi n}}, \tag{22}$$

$$\to 0.$$

By (22), the matching rate (i.e., the average number of matches per unit time) for the greedy policy converges to $n\lambda$ as $n \to \infty$.

When a match occurs (i.e., when there is at least one available mate upon an agent's arrival), the expected number of available mates when an agent arrives can be approximated by

$$\sqrt{n}E[N(0,\lambda/\eta)|N(0,\lambda/\eta) > 0] = \frac{\lambda}{\eta}\sqrt{\frac{2n}{\pi}}. \tag{23}$$

By (22)–(23) and Lemma 1, the utility rate of the greedy policy, which is denoted by U_n^g, satisfies

$$U_n^g \sim n\lambda m\left(\frac{\lambda}{\eta}\sqrt{\frac{2n}{\pi}}\right). \tag{24}$$

5. Simulation Results

To assess the accuracy of our asymptotic results, we consider special cases of the three canonical examples in Section 10 of the online appendix: exp(1), Pareto(1, 2), and U[0, 1]. For all cases, we let $\lambda = \eta = 1$ and $n =$ 1,000, so that the mean number of buyers and sellers in a match-free system is 1,000. We initialize the system with 1,000 buyers and 1,000 sellers, simulate the system for 1,500 time units, discarding the first 150 time units, and then repeat this procedure 100 times.

To find the optimal population threshold levels, we compute the utility rate for the population threshold policy for each integer threshold value in the range [0, 1,000], using the same set of random numbers for each threshold level. We repeat the same procedure for the utility threshold policy and discretize the utility threshold values by 0.1 for the exp(1) and Pareto(1, 2) cases and by 0.01 for the U[0, 1] case.

5.1. Exponential(1) Case

In the exponential case, we predict that the optimal population threshold level is $z_n^* = \frac{1,000}{\ln 1,000} = 144.8$, and the utility rate under this threshold policy approaches the upper bound and is twice as large as the utility rate of the greedy policy (see Section 10.1 of the online appendix). The optimal threshold level found via simulation is 148, and the suboptimality of the utility rate under the threshold 144.8 versus the threshold 148 is 0.004% (Table 2). Our heuristic utility threshold is $v_n^* =$ 5.56 from (123) in the online appendix, which coincides with the optimal threshold found via simulation (with a discretization of 0.1) of 5.6.

However, the predicted utility rates are less accurate than our determination of the best threshold levels. By (119) in the online appendix, our best estimate for the utility rate under the optimal population threshold policy is 5,553, which is 14.9% higher than the simulated value in Table 2. By (21) and (117) in the online appendix, our best estimate of the utility rate under the greedy policy is

$$U_n^g \approx \frac{\lambda n}{\nu}\left(1 - \frac{1}{\sqrt{2\pi n}}\right)\left(\gamma + \ln\left(\frac{\lambda}{\eta}\sqrt{\frac{2n}{\pi}}\right)\right),$$

$$= 3,757,$$

which is 8.5% higher than the simulated value in Table 2. Our best estimate of the upper bound is given in (115) in the online appendix, which yields 7,485. The optimal-to-greedy ratio of the simulated utility rates is $\frac{4,833}{3,462} = 1.40$ rather than 2. Further simulations reveal that convergence is very slow: this simulated ratio is 1.48 when $n = 10^4$ and 1.54 when $n = 10^5$. Most of the inaccuracy in estimating the optimal-to-greedy ratio is due to the fact that the simulated utility rate of the optimal threshold policy is not very close to the upper bound.

Table 2. Theoretical and Simulation Results for the Population Threshold Policy

	Optimal population threshold		Simulated utility rate [95% confidence interval]		
Utility distribution	Theoretical	Simulation	Theoretical threshold	Simulation threshold	Greedy policy
Exponential(1)	144.8	148	4,833 [4,824, 4,840]	4,833 [4,827, 4,841]	3,462 [3,425, 3,503]
Pareto(1, 2)	333.3	347	22,095 [21,997, 22,241]	22,102 [21,972, 22,234]	8,259 [8,107, 8,428]
Uniform(0, 1)	0	22	908.4 [906.0, 911.3]	946.3 [945.1, 947.7]	908.4 [906.0, 911.3]

Table 3. Simulation Results for Both Threshold Policies

	Population threshold policy			Utility threshold policy		
Utility distribution	Optimal threshold	Utility rate	Fraction abandoned	Optimal threshold	Utility rate	Fraction abandoned
Exponential(1)	148	4,833 [4,827, 4,841]	0.140	5.6	5,732 [5,724, 5,740]	0.150
Pareto(1, 2)	347	22,102 [21,972, 22,234]	0.334	42.0	43,750 [43,541, 43,960]	0.503
Uniform(0, 1)	22	946.3 [945.1, 947.7]	0.027	0.96	963.0 [961.7, 964.2]	0.021

Note. Columns 2 and 3 are taken from Table 2.

Finally, the utility rate of the optimal utility threshold policy is 5,732 (Table 3). Although still far from the upper bound, it is 18.6% higher than the utility rate achieved by the optimal population threshold policy.

5.2. Pareto(1, 2) Case

In the Pareto case, we predict that the optimal population threshold level is $\frac{1,000}{3} = 333.3$. The optimal population threshold found via simulation is 347, and the utility suboptimality of the theoretical threshold is 0.03% (Table 2). The solution to (130) in the online appendix is $z^* = 0.512$. Hence, the optimal utility threshold level in (131) in the online appendix is $v_n^* = 42.8$, which is very close to the value of 42.0 found via simulation.

Our estimate of the utility rate under the optimal population threshold policy is 21,573 by (127) in the online appendix, which is 2.4% less than the simulated value of 22,102 in Table 2. The utility rate under the optimal utility threshold policy in (132) in the online appendix is 43,756, which is nearly identical to the optimal simulated value of 43,750. Our best estimate for the utility rate of the greedy policy is $(1 - \frac{1}{\sqrt{2\pi n}})$ times the right-hand side of (125) in the online appendix, or 8,791, which is 6.4% larger than the simulated value in Table 2. Our estimate of the upper bound in (124) in the online appendix is 56,050. By (129) in the online appendix, the predicted performance ratio between the optimal population threshold policy and the greedy policy is $\frac{2}{3}(\frac{1,000\pi}{18})^{1/4} = 2.42$, compared with the optimal-to-greedy simulated ratio of $\frac{22,102}{8259} = 2.68$ (Table 2). By (124) in the online appendix, the ratio of the upper bound to the utility rate of the optimal population threshold policy is predicted to be $\frac{3\sqrt{3}}{2} = 2.60$, compared with the simulated value of $\frac{56,050}{22,102} = 2.54$.

The simulated utility rate of the optimal utility threshold policy is nearly twice as large as the simulated utility rate of the optimal population threshold policy (Table 3), although it is still 21.9% smaller than the predicted upper bound of 56,050.

5.3. Uniform(0, 1) Case

In the uniform case, we predict that the greedy policy is asymptotically optimal. The optimal population threshold level found via simulation is 22, and the resulting utility suboptimality of the greedy policy is 4.0% (Table 2). Note that other population thresholds aside from zero are also asymptotically optimal in this case, including $\ln(n) = \ln(1,000) = 6.91$, which has a suboptimality of 2.0%. Our best estimate of the utility rate under the greedy policy is $(1 - \frac{1}{\sqrt{2\pi n}})$ times the right-hand side of (135) in the online appendix, or 948.2, which is 4.4% larger than the simulated value in Table 2. The upper bound in (133) in the online appendix equals 987.4, which is 4.3% larger than the utility rate corresponding to the optimal population threshold level of 22. The predicted optimal utility threshold from Equation (137) in the online appendix is $v_n^* = 0.974$, compared with the value of 0.96 found via simulation, for a utility suboptimality of 0.24% (Table 3).

In summary, our analysis identifies the optimal threshold level within about 2% (considering the possible range of [0, 1,000]) and its suboptimality is no more than 2% for the population threshold policy in the uniform case and is negligible in the other five cases. We also note that the predicted fraction of agents who abandon the market under the optimal population threshold policy, which is $\frac{2z_n}{2\lambda_n} = \frac{1}{\ln n} = 0.145$ (i.e., the total abandonment rate divided by the total arrival rate) in the exponential case, $\frac{z_n^*}{n} = \frac{1}{3}$ in the Pareto case by (126) in the online appendix, and $1 - \frac{1}{\sqrt{2\pi n}} = 0.013$ in the uniform case by (21), are reasonably close to the simulated values in the fourth column of Table 3. As predicted by our analysis, the utility rate of the greedy policy—normalized by the mean of the matching distribution—increases with the right tail of the matching distribution (this quantity is 1,817 for the uniform, 3,462 for the exponential, and 4,129 for the Pareto), as does the ratio of the utility rates between the best threshold policy and the greedy policy (1.04 for the uniform, 1.40 for the exponential, and 2.68 for the Pareto under the population threshold policy, and 1.06, 1.66, and 5.30 under the utility threshold policy). In addition, despite the asymptotic optimality result, there is a large gap between the utility rate of the best population threshold policy and

the upper bound in the exponential case. The improvement of the utility threshold policy over the population threshold policy also increases with the right tail of the matching distribution, with the ratio of the utility rates equaling 1.02, 1.19, and 1.98 for the uniform, exponential, and Pareto cases, respectively. This improvement is achieved by being more patient and allowing more agents to abandon the market, particularly in the Pareto case (last column in Table 3).

6. Unbalanced Markets

In this section, we consider unbalanced markets, where buyers and sellers arrive at rates $n\lambda_b$ and $n\lambda_s$ in the nth system, and abandon at rates η_b and η_s, respectively. We restrict ourselves to the analysis of the utility threshold policy in the case $\alpha \in (0,1)$, which is very similar to the corresponding analysis in the symmetric case. We also note that an analysis of the population threshold policy in the unbalanced case is complicated by the extra degree of freedom that is introduced (and needs to be determined) in Equations (45)–(47) in the online appendix and is beyond the scope of this paper. Under the utility threshold policy in the nth system, an arriving buyer is matched to the seller that yields the maximum utility if this utility exceeds the threshold $v_{n,s}$; similarly, an arriving seller is matched to its highest-matching buyer if the utility exceeds the threshold $v_{n,b}$. The dynamics are given by the equations

$$\begin{aligned} B_n(t) = B_n(0) &+ \sum_{j=1}^{N_B^+(n\lambda_b t)} I_{\left\{\max_{i=1}^{S_n\left(A_{j-}^B\right)} V_{i,j}^B \le v_{n,s}\right\}} \\ &- \sum_{j=1}^{N_S^+(n\lambda_s t)} I_{\left\{\max_{i=1}^{B_n(A_{j-}^S)} V_{i,j}^S > v_{n,b}\right\}} \\ &- N_B^-\left(\eta_b \int_0^t B_n(r_-)dr\right), \end{aligned} \tag{25}$$

$$\begin{aligned} S_n(t) = S_n(0) &+ \sum_{j=1}^{N_S^+(n\lambda_s t)} I_{\left\{\max_{i=1}^{B_n(A_{j-}^S)} V_{i,j}^S \le v_{n,b}\right\}} \\ &- \sum_{j=1}^{N_B^+(n\lambda_b t)} I_{\left\{\max_{i=1}^{S_n(A_{j-}^B)} V_{i,j}^B > v_{n,s}\right\}} \\ &- N_S^-\left(\eta_s \int_0^t S_n(r)dr\right), \end{aligned} \tag{26}$$

where $\{A_j^B : j \ge 1\}$ is the sequence of arrival times associated with $N_B^+(n\lambda_b \cdot)$, and $\{A_j^S : j \ge 1\}$ is the sequence of arrival times associated with $N_S^+(n\lambda_s \cdot)$. As in the symmetric case, the $V_{i,j}^B$'s and $V_{i,j}^S$'s are independent arrays of i.i.d. random variables with distribution $F(\cdot)$.

Following the development in the symmetric case (e.g., (16)), the utility rate takes the form

$$\begin{aligned} U_n^u(v_{n,b}, v_{n,s}) = n\lambda_b E[E[M(S_n(\infty)) I_{\{M(S_n(\infty)) \ge v_{n,s}\}} | S_n(\infty)]] \\ + n\lambda_s E[E[M(B_n(\infty)) I_{\{M(B_n(\infty)) \ge v_{n,b}\}} | B_n(\infty)]]. \end{aligned} \tag{27}$$

Our main result is presented in Theorem 3. The proof of Theorem 3 appears in Section 9.4 in the online appendix and largely mimics the proof of Theorem 2.

Theorem 3. *Suppose that Assumption 3 holds. Let v_* be the optimal solution for the optimization problem*

$$\max_{v \ge 0} \lambda_s b^\alpha E[X I_{\{b^\alpha X > v\}}] + \lambda_b s^\alpha E[X I_{\{s^\alpha X > v\}}] \tag{28}$$

$$\text{subject to } \eta_b \mathrm{b} + \lambda_s = \lambda_b \exp\left(-\kappa \mathrm{s}/\mathrm{v}^{1/\alpha}\right) + \lambda_s \exp\left(-\kappa \mathrm{b}/\mathrm{v}^{1/\alpha}\right), \tag{29}$$

$$\eta_b b + \lambda_s = \eta_s s + \lambda_b. \tag{30}$$

Then a threshold policy of the form $v_{n,b}^* = v_{n,s}^* = m(n)v_*$ is asymptotically optimal among the class of utility threshold polices. The associated utility rate satisfies

$$\begin{aligned} \lim_{n\to\infty} \frac{U_n^u(v_{n,b}^*, v_{n,s}^*)}{nm(n)} = \lambda_s {b_*}^\alpha E[X I_{\{b_*^\alpha X > v_*\}}] \\ + \lambda_b s_*^\alpha E[X I_{\{s_*^\alpha X > v_*\}}], \end{aligned}$$

where b_*, s_* are solutions that satisfy constraints (29)–(30).

Although the results in Theorem 3 are beyond our intuitive grasp, we attempt to provide some possible intuition for why $v_b^* = v_s^*$ in the unbalanced case. Let us consider a fluid model in which the two utility thresholds are both equal to v^*. We can classify the matched sellers into two categories: actively matched (i.e., they arrive to the market and are immediately matched with buyers) and passively matched (i.e., they wait in the market and then are matched with arriving buyers). Now suppose that we change the utility thresholds to v_b and v_s, where $v_b < v^* < v_s$, in such a way that the number of additional actively-matched sellers (call it ds) equals the reduction in the number of passively-matched sellers. Because the total number of matched sellers does not change, the number of abandoned agents remains the same and we can focus on the matching utilities of these marginal sellers. Let u_a be the utility per match for the ds additional actively-matched sellers, and let u_p be the utility rate per match for the ds sellers that are no longer passively matched. The utility per match of these marginal sellers is between the new threshold and the old threshold, and therefore $v_b < u_a < v* $ and $v* < u_p < v_s$.

Hence, the net change in utility is $(u_a - u_p)ds$, which is negative.

We conclude this section with a numerical example that is a variant of the one in Section 10.2 of the online appendix: let $\lambda_b = 2$, $\lambda_s = 1$, $\eta_b = 1$, $\eta_s = 1$, $n = 1{,}000$, and assume a Pareto(1, 2) distribution, so that $\alpha = 1/2$, $\kappa = 1/\pi$ and $m(n) = \sqrt{1{,}000\pi}$. Then $b(s) = s + 1$ in (96) in the online appendix, and (105) in the online appendix reduces to

$$2e^{-s\tau} + e^{-(s+1)\tau} = s + 2.$$

The solution to (98) in the online appendix is $s_* = 0.365$ and $\tau(0.365) = 0.361$, which yields $v^*_{n,b} = v^*_{n,s} = \sqrt{\frac{1{,}000}{0.361}} = 52.7$. Interestingly, this threshold level of 52.7 is higher than in the symmetric case, where $\lambda_b = 1$ and $v^*_n = 42.8$. Moreover, leaving all parameter values fixed except for λ_b, we numerically compute $v^*_{n,b}$ in (104) in the online appendix and find that it is increasing and concave in $\lambda_b \geq 1$.

With $\lambda_b = 2$, we simulate this system in the same manner as in Section 5. At a discretization of 0.1, a two-dimensional search of $(v_{n,b}, v_{n,s})$ space via simulation for the optimal thresholds yields (52.7, 52.3), with a corresponding simulated utility rate of 71,046 and with abandonment fractions of 0.681 for buyers and 0.363 for sellers. The simulated utility rate at $(v^*_{n,b}, v^*_{n,s}) = (52.7, 52.7)$ is 71,010, which is suboptimal by 0.05%. The predicted utility rate, $U^u_n(v^*_{n,b}, v^*_{n,s})$ in (106) in the online appendix, is 70,992, which is 0.03% less than the simulated value of 71,010.

Fixing one utility threshold level at 52.7 and varying the other threshold level (figure 1 in the online appendix) reveals that the simulated utility rate is slightly more sensitive to v_s than v_b, perhaps because arriving buyers see fewer potential matches than arriving sellers. This figure also shows that it is more suboptimal to underestimate the threshold level than to overestimate it.

7. Batch-and-Match Policy

In this section, we restrict ourselves to Pareto matching utilities with finite mean, where $F(v) = 1 - (cv)^{-\beta}$, for $\beta > 1$, $c > 0$, and $cv \geq 1$, so that $\alpha = 1/\beta$ in Assumption 2. We consider a one-parameter batch-and-match policy: At times $t = \{\Delta, 2\Delta, 3\Delta, \ldots,\}$, we match $\min\{B(t), S(t)\}$ buyers and sellers by randomly choosing $\min\{B(t), S(t)\}$ agents from the thicker side of the market (e.g., buyers if $B(t) \geq S(t)$) and then maximize the total utility from these matches; this class of policies allows us to consider a balanced random assignment problem, which is easier to analyze than an unbalanced random assignment policy. The goal is to choose the time window Δ that maximizes the long-run average utility rate. The main qualitative conclusion from this section is that—for the special case of $\lambda = \eta = c = 1$ and $\beta = 2$—the utility threshold policy easily outperforms this batch-and-match policy, both in the asymptotic analysis and in the simulation results.

To analyze the performance of this policy, we consider a random assignment problem, where there are k buyers and k sellers with i.i.d. Pareto matching utilities $V_{i,j}$ between buyer i and seller j. The matching problem is

$$\max_{\pi} \sum_{i=1}^{k} V_{i,\pi(i)},$$

where π is a permutation function. Let $\mathcal{M}(k) = \max_\pi \sum_{i=1}^k V_{i,\pi(i)}$.

The main result of this section is given in Theorem 4, which is proved in Section 9.5 of the online appendix. The corresponding results for the unbalanced case are presented without proof at the end of Section 9.5 of the online appendix.

Theorem 4. *Consider the symmetric model with arrival rate λ, abandonment rate η, and Pareto (c, β) matching utilities with finite mean. Let $U^b_n(\Delta)$ be the utility rate of the batch-and-match policy with time window Δ. Then the utility rate satisfies*

$$\lim_{n\to\infty} \frac{U^b_n(\Delta)}{n^{\alpha+1}} \leq \frac{c\Gamma(1-\alpha)\left[\frac{\lambda}{\eta}(1 - e^{-\eta\Delta^*})\right]^{\alpha+1}}{\Delta^*}, \tag{31}$$

where the asymptotically optimal time window Δ^* is the unique solution to

$$e^{\eta\Delta} = (1 + \alpha)\eta\Delta + 1. \tag{32}$$

Note that Δ^* increases in α in (32), and so—as in the population threshold policy—heavier tails lead to thicker markets.

We conclude this section with the numerical Pareto(1, 2) example from Section 5, where $\lambda = \eta = 1$, $n = 1{,}000$, $F(v) = 1 - v^{-2}$, and $\alpha = 1/2$. Equation (32) reduces to $e^\Delta = \frac{3}{2}\Delta + 1$, which has solution $\Delta^* = 0.76$, implying from (112) in the online appendix that we make approximately $1{,}000(1 - e^{-0.76}) = 532$ matches in each cycle. The upper bound for the utility rate in (31) is $\frac{\sqrt{\pi}(1-e^{-\Delta^*})^{3/2}1{,}000^{3/2}}{\Delta^*} = 28{,}644$, which is much smaller than the predicted utility rate of 43,750 for the utility threshold policy from (132) in the online appendix. A one-dimensional search using simulation generates an optimal time window of 0.75, confirming the accuracy of our asymptotic analysis. The simulated utility rate under this time window is 25,168 and the lower bound for the utility rate in Lemma 5 is 131, suggesting that the upper bound is useful and the lower bound is very loose.

8. Concluding Remarks

A fundamental trade-off in centralized dynamic matching markets relates to market thickness: whether matches should be delayed—at the risk of antagonizing waiting agents—in the hope of obtaining better matches in the future. Very little is known about this issue when matching utilities are general. By combining queueing asymptotics (as an aside, we note that perhaps the most surprising part of our study is that rather than requiring a diffusion analysis, a fluid analysis is sufficient to analyze this problem) with extreme value theory, we obtain explicit results that shed light on this issue. For symmetric markets, as the right tail of the matching utility distribution gets heavier, it is optimal to become more patient and let the market thickness (and abandonment rate) increase. Whereas empirical works on matching markets use more complicated covariate models than what we consider (e.g., Hitsch et al. 2010, Boyd et al. 2013, Agarwal 2015), it seems clear from these analyses that matching utilities typically are not in the domain of attraction of the Weibull law. Therefore, large centralized matching markets—whether balanced or unbalanced—are likely to benefit from allowing the market to thicken.

Enabled by the decoupling of the fluid queueing dynamics and the extremal behavior of the matching utilities, our study appears to be the first to allow for correlated matching utilities, which is likely to be a common phenomenon in practice: an agent who is deemed objectively attractive in a labor, housing, or school choice model is likely to have matching utilities with potential mates that are positively correlated rather than i.i.d. In Section 10.4 of the online appendix, we find that positive correlation reduces the market thickness in the utility threshold policy but not the population threshold policy, and it reduces the utility rate under both policies.

We note four limitations in our study. First, most of our analysis is restricted to arrival-only policies. In particular, it might be possible to do better by batching sets of agents and then matching them, as in Mertikopoulos et al. (2020). Moreover, generalizing their results to our setting is likely to be quite challenging, in that the $\pi^2/6$ result requires an exponential matching distribution and an objective of minimizing the matching cost (they minimize mismatch plus waiting costs rather than maximizing utility in the presence of abandonment). Whereas they generalize their results in section 6 of their paper by positing a functional form for how the expected minimum mismatch costs decrease as a function of the number of agents in the market, this functional form does not appear to follow from any more primitive distributional assumptions. In Section 7, we consider Pareto utilities and analyze a simple batch-and-match policy, which periodically (with an asymptotically optimal time window) optimally matches all agents on the thinner side of the market with an equal number of agents randomly selected from the thicker side of the market. Perhaps surprisingly, we show that, in the Pareto case, the utility threshold policy easily outperforms the batch-and-match policy. Nonetheless, this does not preclude the possibility that more sophisticated batching policies (e.g., optimally—rather than randomly—select the agents to match from the thicker side of the market or include a utility threshold for allowable matches as a second parameter) might outperform the utility threshold policy.

Second, most of our analysis considers a symmetric market, with buyers and sellers having the same arrival and abandonment rates. Whereas some markets, such as cadaveric organ transplants and public housing, tend to have chronic supply shortages, other markets have economic forces at play that tend to roughly balance supply and demand. In a static matching market, even a slight imbalance can give rise to a unique stable matching (Ashlagi et al. 2017a). We also note that a greedy policy is optimal in a somewhat different unbalanced market setting, where easy-to-match agents can match with all other agents in the market with a specified probability, but hard-to-match agents can match only with easy-to-match agents with a different specified probability (Ashlagi et al. 2018a). In our analysis of the utility threshold policy in the heavy-tailed case of an unbalanced market, we obtain the somewhat surprising result that the solution is symmetric; that is, the utility threshold is the same for buyers and sellers. Moreover, we find (in our Pareto example) that the amount of patience increases with the amount of imbalance; that is, the larger the imbalance, the more agents that are going to be turned away, and the more selective the matching becomes. However, we leave a complete analysis of the unbalanced problem for future work.

Although our model can be viewed as allowing a continuum of classes via the distribution of the matching utility, the third restriction is that our analysis does not naturally lend itself to a setting where there is a discrete number of classes with class-dependent matching utilities. In particular, in some settings (e.g., organ donation), some classes of buyers/sellers are compatible with only prespecified classes of sellers/buyers. The decoupling of the queueing fluid dynamics and the extremal behavior of the matching utilities should carry over to the setting with a finite number of classes with some incompatibility among classes. However, it would make sense to consider multiple thresholds in this setting, and a multidimensional model with multiple thresholds would be a nontrivial extension.

The final restriction is exponential abandonment. Relaxing this assumption would require a different approach, such as hazard rate scaling (Reed and Tezcan 2012) and would likely be much more difficult.

Finally, we note that there may be equity issues if a significant number of agents are allowed to abandon the market (Table 3). The consideration of a risk-sensitive objective function would likely require a diffusion approximation, which would be, for example, a two-dimensional Ornstein-Uhlenbeck process with an unusual Skorokhod condition under a population threshold policy.

Acknowledgments

The authors thank Can Wang and Halwest Mohammad for running some of the simulations described in Sections 5 and 6 and Kavita Ramanan for advice about the Skorokhod mapping in Section 9 of the online appendix.

References

Adan I, Weiss G (2012) Exact FCFS matching rates for two infinite multitype sequences. *Oper. Res.* 60(2):475–489.

Agarwal N (2015) An empirical model of the medical match. *Amer. Econom. Rev.* 105(7):1939–1978.

Akbarpour M, Li S, Oveis Gharan S (2020) Thickness and information in dynamic matching markets. *J. Political Econom.* 128(3): 783–815.

Aldous DJ (2001) The $\zeta(2)$ limit in the random assignment problem. *Random Structures Algorithms* 18(4):381–418.

Anderson R, Ashlagi I, Kanoria Y, Gamarnik D (2017) Efficient dynamic bargain exchange. *Oper. Res.* 65(6):1446–1459.

Ashlagi I, Jaillet P, Manshadi PH (2013) Kidney exchange in dynamic sparse heterogenous pools. Preprint, submitted January 15, https://arxiv.org/abs/1301.3509.

Ashlagi I, Kanoria Y, Leshno JD (2017a) Unbalanced random matching markets: The stark effect of competition. *J. Political Econom.* 125(1):69–98.

Ashlagi I, Nikzad A, Strack P (2018a) Matching in dynamic imbalanced markets. Preprint, submitted September 18, https://arxiv.org/abs/1809.06824.

Ashlagi I, Burq M, Jaillet P, Manshadi V (2019) On matching and thickness in heterogeneous dynamic markets. *Oper. Res.* 67(4): 927–949.

Ashlagi I, Burq M, Dutt C, Jaillet P, Saberi A, Sholley C (2018b) Maximum weight online matching with deadlines. Preprint, submitted August 9, https://arxiv.org/abs/1808.03526.

Ashlagi I, Azar Y, Charikar M, Chiplunkar A, Geri O, Kaplan H, Makhijani R, et al. (2017b) Min-cost bipartite perfect matching with delays. Jansen K, Rolim JDP, Williamson D, Vempala SS, eds. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (Schloss Dagstuhl–Leibniz-Zentrum fur Informatik, Dagstuhl, Germany), 1–20.

Baccara M, Lee S, Yariv L (2020) Optimal dynamic matching. *Theoret. Econom.* 15(3):1221–1278.

Bojanic R, Seneta E (1971) Slowly varying functions and asymptotic relations. *J. Math. Anal. Appl.* 34(2):302–315.

Boyd D, Lankford H, Loeb S, Wyckoff J (2013) Analyzing the determinants of the matching of public school teachers to jobs: Estimating compensating differentials in imperfect labor markets. *J. Labor Econom.* 31(1):83–117.

Büke B, Chen H (2017) Fluid and diffusion approximations of probabilistic matching systems. *Queueing Systems* 86:1–33.

Bušić A, Meyn S (2015) Approximate optimality with bounded regret in dynamic matching models. *ACM Sigmetrics Performance Evaluation Rev.* 43(2):75–77.

Caldentey R, Kaplan EH, Weiss G (2009) FCFS infinite bipartite matching of servers and customers. *Adv. Appl. Probab.* 41(3):695–730.

Ding Y, McCormick ST, Nagarajan M (2021) A fluid model for one-sided bipartite queues with match-dependent rewards. *Oper. Res.* 69(4):1256–1281.

Duffie D, Qiao L, Sun Y (2018) Dynamic directed random matching. *J. Econom. Theory* 174:124–183.

Emek Y, Kutten S, Wattenhofer R (2016) Online matching: Haste makes waste! *Proc. 48th Annual ACM SIGACT Sympos. Theory Comput.* (ACM, New York), 333–344.

Ethier SN, Kurtz TG (2005) *Markov Processes: Characterization and Convergence* (John Wiley & Sons, New York).

Galambos J (1978) *The Asymptotic Theory of Extreme Order Statistics* (Wiley, New York).

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

Gumbel EJ (1958) *Statistics of Extremes* (Columbia University Press, New York).

Gurvich I, Ward A (2014) On the dynamic control of matching queues. *Stochastic Systems* 4(2):479–523.

Hitsch GJ, Hortacsu A, Ariely D (2010) Matching and sorting in online dating. *Amer. Econom. Rev.* 100(1):130–163.

Hu M, Zhou Y (2021) Dynamic type matching. *Manufacturing Service Oper. Management*, ePub ahead of print March 18, https://doi.org/10.1287/msom.2020.0952.

Kaplan EH (1988) A public housing queue with reneging and task-specific servers. *Decision Sci.* 19(2):383–391.

Karp RM, Vazirani UV, Vazirani VV (1990) An optimal algorithm for on-line bipartite matching. *Proc. 22nd Annual ACM Sympos. Theory Comput.* (ACM, New York), 352–358.

Liu X, Gong Q, Kulkarni VG (2015) Diffusion models for double-ended queues with renewal arrival processes. *Stochastic Systems* 5(1):1–61.

Mertikopoulos P, Nax HH, Pradelski BSR (2020) Quick or cheap? Breaking points in dynamic markets. Preprint, submitted January 20, https://arxiv.org/abs/2001.00468.

Mezard M, Parisi G (1987) On the solution of the random link matching problems. *J. Phys.* 48(9):1451–1459.

Moyal P, Perry O (2017) On the instability of matching queues. *Ann. Appl. Probab.* 27(6):3385–3434.

Nazari M, Stolyar AL (2019) Reward maximization in general dynamic matching systems. *Queueing Systems* 91:143–170.

Reed J, Tezcan T (2012) Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Oper. Res.* 60(4):981–995.

Resnick S (1987) *Extreme Values, Regular Variation, and Point Processes* (Springer, New York).

Ünver MU (2010) Dynamic kidney exchange. *Rev. Econom. Stud.* 77(1):372–414.

Varma SM, Bumpensanti P, Maguluri ST, Wang H (2019) Dynamic pricing and matching in two-sided queues. Preprint, submitted November 6, https://arxiv.org/abs/1911.02213.

Jose H. Blanchet is a professor in the Department of Management Science and Engineering at Stanford University. He serves on the editorial board of *Mathematics of Operations Research*, *Stochastic Systems*, *Insurance: Mathematics and Economics*, and *Extremes*.

Martin I. Reiman is a professor in the Industrial Engineering and Operations Research Department at Columbia University. Prior to this, he was a Distinguished Member of Technical Staff at Bell Labs in Murray Hill, New Jersey. His

research is focused on the analysis, design, and control of stochastic service systems, primarily using asymptotics.

Virag Shah is a data scientist at Uber, Inc. Prior to this, he was a postdoctoral scholar at the Management Science and Engineering Department at Stanford University. He has also held research positions at the Microsoft Research-INRIA Joint Center in Paris, the University of Texas at Austin, and the Indian Institute of Science at Bangalore. His research interests include pricing, matching, and experimentation in online marketplaces.

Lawrence M. Wein is the Jeffrey S. Skoll Professor of Management Science and a Senior Associate Dean of Academic Affairs at Stanford University's Graduate School of Business. His primary research interest is studying public sector problems. He is a former editor-in-chief of *Operations Research*.

Linjia Wu is a PhD student in the Department of Management Science and Engineering at Stanford University. Her research interests are mainly in matching and experimental design for online platforms. She received her undergraduate degree at Peking University.