Public Housing Assignment in Pittsburgh: A Case Study

Yichuan Ding, S. Thomas McCormick, Mahesh Nagarajan
Sauder School of Business, University of British Columbia, Vancouver, BC V6T1Z2,
{Daniel.Ding, Tom.McCormick, Mahesh.Nagarajan}@sauder.ubc.ca

This case study applies the machinery developed in (Ding et al. 2018) to a practical problem – public housing allocation in the city of Pittsburgh. Specifically, Ding et al. (2018) developed a fluid model that approximates a bipartite queueing system (BQS) equipped with a Matching(M)+Waiting(W) index, which is suitable for modeling the public housing allocation system. In particular, we can calculate the fluid process over any finite interval using Algorithm 4 in (Ding et al. 2018), which predicts how the queue-lengths in the BQS evolve in a given time window. We aim to illustrate the following points via this case study.

1. The prediction from the fluid model provides accurate approximations to the stochastic system.
2. By identifying the routing components and the associated switch times, our machinery helps the system manger better understand the allocation outcome under a certain policy.
3. Our machinery enables the policy designer to evaluate and compare system performances under different M+W indices.

1. Overview

Most major North American cities provide affordable housing options for eligible low-income families, seniors, and persons with disabilities[1]. The housing authority (HA) in each city is responsible for the maintenance and assignment of such houses. Currently, such programs accommodate approximately 1.2 million households[2], but the supply is far from large enough to meet all the demand for housing. Applicants to such programs typically have strict constraints on the housing options they prefer based on location, the community type, and features of the housing unit such as

[1] The rent for public housing is usually around 30% of the renter's gross income. This can vary depending on the size of the residence and the type of the community.

[2] `http://portal.hud.gov/hudportal/HUD?src=/topics/rental_assistance/phprog`

the number of rooms etc., which exacerbates the shortage for some types of housing. For example, the waiting time for single-bedroom apartments was reported to be around 28 years in the District of Columbia (Stockton 2013).

We illustrate the use of the tools we developed in this paper using data from the Housing Authority of the City of Pittsburgh (HACP) that was gathered by Geyer and Sieg (2013). The HACP requires applicants to specify their preference among the available types of units, and among the communities where public housing is available. Houses are allotted based on availability in the preferred community and according to the applicants' registration dates when joining the waitlist for housing. Therefore the waitlist can be effectively partitioned into multiple queues according to the applicants' selections. If we assume that the priority is first-come-first-served in each queue[3], then the waitlist system is a BQS.

We take the following steps to study the optimal scoring policies that the HACP could use to allocate their houses. First, Section 2 estimates parameters of this BQS using historical data that was obtained from the HACP and the data from the Survey of Income and Program Participation (SIPP). Next, Section 3 calculates the fluid limit process for this BQS using the tools developed in this paper. We compare the fluid limit process to the stochastic process that is simulated using the same set of parameters. The comparison shows that the fluid model generates very robust predictions. Lastly, Section 4 shows that the predictions of the our fluid model make it easy to evaluate different scoring rules with respect to various performance measures.

2. Parameter Estimation

The average net number of households that move out of PH1, PH2, and PH3 per quarter are estimated by Geyer and Sieg (2013), which can be regarded as the resource arrival rates in the BQS mode, namely $\mu_1 = 118.9$, $\mu_2 = 29.5$, and $\mu_3 = 9.1$.

Geyer and Sieg (2013) developed a discrete-choice-model to predict how low-income households select their housing community by studying the public housing allocation system of HACP. Although our objective in this paper is different than theirs, we use data sets from their study and calibrate many useful system parameters based on estimates reported in their study. They identified 34 community types in the city of Pittsburgh, and further categorized them into six broad community types, each of which consists of fairly homogeneous housing units. To make our BQS even simpler, we consider only three of these six broad community types, denoted by PH1, PH2, and PH3. The majority of the housing units in these three broad types are offered to non-senior families, so we can focus on these three types independently from the other types.

[3] In fact, most households with the same selection of community types will be sequenced according to their registration date. The HACP may prioritize a small number of households with urgent needs, but this is often negligible.

Applicants specify a first choice among the three types, and are encouraged (but not required) to add a second choice. An applicant's choice profile can then be represented by a single number or an ordered pair. For example, choice profile [12] means that the applicant's first choice is PH1 and second choice is PH2, and choice profile [2] means that the applicant's first choice is PH2 and they did not specify a second choice.

To estimate the proportion of each choice profile, we use the discrete-choice model estimated by Geyer and Sieg (2013), which assumes a renter's utility to be

$$\begin{aligned} u_{jt} &= \gamma_j + \beta \ln(y_{jt}) + \delta x_t + \mathrm{mc} + \xi_{jt},\ j = 1,2,3 \\ u_{0t} &= \ln(y_{0t}) + \delta x_t + \xi_{0t},\ j = 1,2,3 \end{aligned} \tag{1}$$

where j denotes the community type, ($j = 0$ means living outside public housing), t is the index of an applicant, γ_j, u_{jt} represents the utility for applicant t to accept a unit of type j, γ_j represents the fixed effect of community type j and y_{jt} denotes the net income of the applicant reduced by the rent of housing type j, β denotes the coefficient for the log of net-income (Geyer and Sieg (2013) assumed $\beta = 1$ for the outside housing option), x_t is a vector of a household's attributes, mc stands for the moving cost for the household (which is incurred when the household accept public housing), and ξ_{jt} represents idiosyncratic shocks which are assumed to be i.i.d. and possess extreme-value distributions.

Geyer and Sieg (2013) studied the choice behavior of householders already residing in public housing as well as those outside public housing. In order to do that, they created two population pools to generate x_t and y_t: the residence pool and the waitlist pool. The residence pool consists of households who resided in public housing in the city of Pittsburgh from June 2001 to June 2006. The attributes of such households were recorded in the HACP's dataset. The waitlist pool consists of $14,234$ randomly selected low-income applicants who were eligible for public housing from 13 metropolitan areas with a public housing-household ratio similar to the city of Pittsburgh. Their information is recorded in the SIPP dataset. Geyer and Sieg (2013) argued that these applicants are a good proxy for those on the waitlist for public housing in Pittsburgh. Since our case study focuses on the dynamics on the waitlist, we ignore possible transfers between different community types (which are small). Therefore, it suffices to create the waitlist pool using the same method described in the above paper and then generate x_t and y_t from that waitlist.

The waitlist pool we use is a subset of the dataset used by Geyer and Sieg (2013). Therefore their choice model can be applied to predict household choice in our system. Using x_t and y_t, we can calculate the deterministic parts of the utility function u_{jt} and u_{0t} using the coefficients reported in Geyer and Sieg (2013). By generating the random errors $(\xi_{jt})_{j=0,1,2,3}$ according to their distributions, we can simulate the values of u_{jt} and u_{0t} and calculate the probabilities for each choice profile. For example, $\Pr(1,2) = \Pr(u_{1t} > u_{2t} > 0)$, $\Pr(2) = \Pr(u_{1t} > 0, u_{2t} < 0, u_{3t} < 0)$. Although our choice set has only three types rather than the six types of (Geyer and Sieg 2013), this does not lead to any estimation bias of the choice probabilities due to the "independence of irrelevant alternatives" property of the discrete choice model.

We estimate our matching function $U(\cdot,\cdot)$ using the following reasoning. (1) The function values should have the same scale for different applicants. Thus, we assign $U(j,(j_1,j_2)) = 1$ when $j = j_1$, i.e., when an applicant's first choice was met; and we assign $U(j,(j_1,j_2)) = 0$ if $j \neq j_1, j_2$, that is, if j is not in an applicant's choice set. (2) For each individual applicant, the degree of matching for the 3 options, i.e., the "first choice", "second choice", and "not interested in public housing" have to reflect an individual's degree of interest among the three options. We find that on average, the second largest value among the random utilities $(u_{jt})_{j=0,1,2,3}$ is 0.66 of the largest value after normalizing "not interested in public housing" to be zero. Therefore, we use 0.66 as the value for the degree of matching if an offered housing unit is the second choice of the applicant.

Based on our estimation from the historical data, new applicants arrive at a constant arrival rate of 443 every quarter of a year. The waitlist was closed in May 2015 and re-opened in July 2015. Assuming that the 1578 applicants on the older closed waitlist for non-senior housing needed to re-register again on the new waitlist and that 50% of these applicants did register on the new waitlist in the first two quarters, yields the total arrival rate of $50\% \times 1578 + 443 = 1232$ per quarter in the first two quarters, and thereafter a stable 443 per quarter since the third quarter.

Finally, applicants on a waitlist may abandon the queue before they are assigned a housing unit for various reasons. Based on our conversations with HACP, the most typical reason that an applicant loses eligibility because of an increase in their income. The data shows that the time for an eligible applicant to lose their eligibility has a similar distribution across different queues, and can be approximated by an exponential distribution with a hazard rate of $d = 0.07$ quarters. Therefore the cumulative distribution of abandonment time is given by $F_i(t) = 1 - \exp(-dt)$.

The parameters we estimated are summarized in Table 1.

3. Comparison of the Fluid Approximation and the Stochastic Process

A fluid approximation to the waitlist system assumes that both applicants and housing units arrive according to deterministic processes. Using Algorithm 4 in (Ding et al. 2018), we can calculate the HOL waiting times in the fluid model using the parameters reported in Table 1. To see whether the fluid approximation is an accurate approximation of the underlying stochastic process, we compare the HOL waiting times predicted by the fluid model to that in a simulated stochastic BQS using the same set of parameters and assuming that all arrival processes are Poisson. We simulate the

Table 1 Summary of Estimated Parameters

Choice Profile i		[12]	[13]	[1]	[21]	[23]	[2]	[31]	[32]	[3]
Probability		0.1013	0.0793	0.049	0.1318	0.1939	0.1061	0.0906	0.1703	0.0776
Community Type j			PH1			PH2			PH3	
Service Rates μ_j			118.9			29.5			9.1	
Time Window			Quarter 1 and 2					Later		
Total Arrival Rate λ			1232					443		
Reneging Rate d						0.07				
$U(\cdot,\cdot)$	Queues	[12]	[13]	[1]	[21]	[23]	[2]	[31]	[32]	[3]
	PH1	1	1	1	0.66	-100	-100	0.66	-100	-100
	PH2	0.66	-100	-100	1	1	1	-100	0.66	-100
	PH3	-100	0.66	-100	-100	0.66	-100	1	1	1

HOL waiting times of this stochastic BQS by bootstrapping each applicant's attributes from the waitlist pool and predicting their choice profiles.

We assume that applicants on the waitlist are ranked by the optimal scoring formula at the steady state, i.e., score* (See Equation (70) in (Ding et al. 2018)), where the weight η is set to be 50. According to Corollary 1, this scoring formula optimizes the steady-state performance with respect to the objective function (5), where efficiency (Ef) and fairness (Fa) are defined according to equations (4) and (3), respectively. Figure 1 plots two sets of HOL waiting time curves over a five-year horizon (20 quarters). One set are predictions made by the fluid model, and the other set are simulated sample paths for the stochastic BQS. The two sets of curves stay fairly close and the fluid process captures the behavior of the simulated sample paths quite well. In the first period, each community PHi only serves the queues that indicated PHi as their first choice. However, this routing configuration changes quickly because scores of queues in routing components of PH2 and PH3 increase at a higher speed than those of other queues due to the acute shortage of housing units in PH2 and PH3. Consequently, a subset of queues served by PH2 and PH3 catch up with queues served by PH1 in their scores, and get merged into the component of PH1. At T_3, queues [32] and [3] in the faster component are merged into the slower component of PH1, which leads to the breakpoints in the waiting-time curves. After T_4, all queues are merged into a single component and this configuration does not further change. We plot the curves for five years, which covers all switch points, though it takes fifteen years in total for the fluid process to reach the steady state. The change in waiting times is smaller than .00001 after 50 years, indicating convergence. The routing configurations in different time periods are depicted in Figure 3.

We repeat the simulation 100 times and calculate the mean absolute deviation (MAD) for the HOL waiting times in each queue, and find MADs to range from 0.097 to 0.116 quarters, which is 8–10 days, which means that the fluid approximation is accurate up to a ten-day period. We also observe that the fluid approximation remains accurate even for queues with thin service rates, such as queues [32] and [3], whose throughput rate is less than 10 per quarter. This demonstrates that fluid models for overloaded systems were quite robust in our experiments.

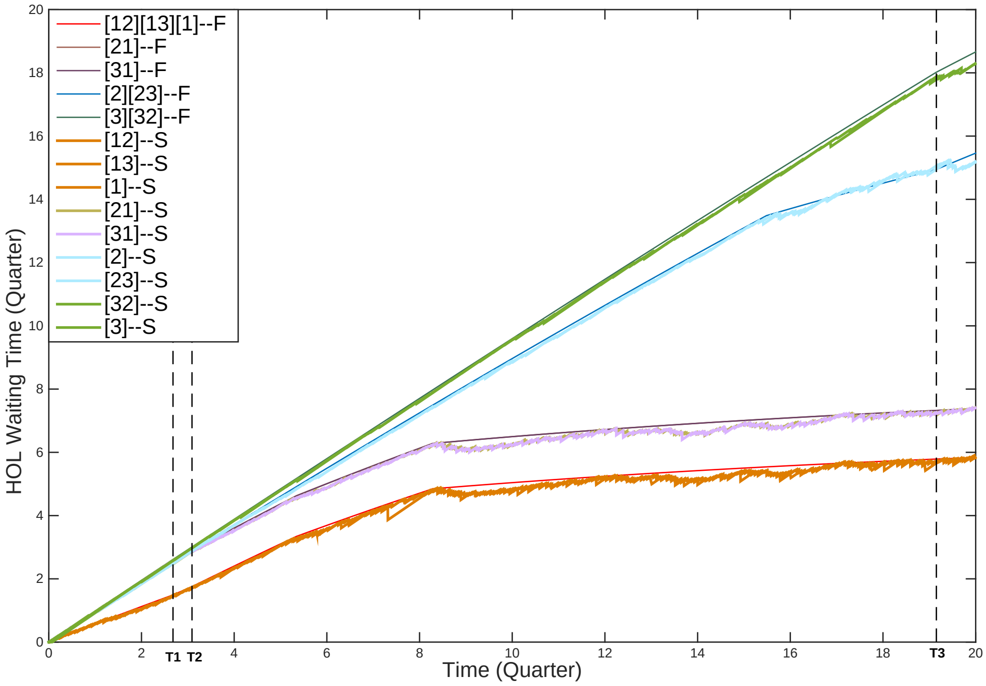


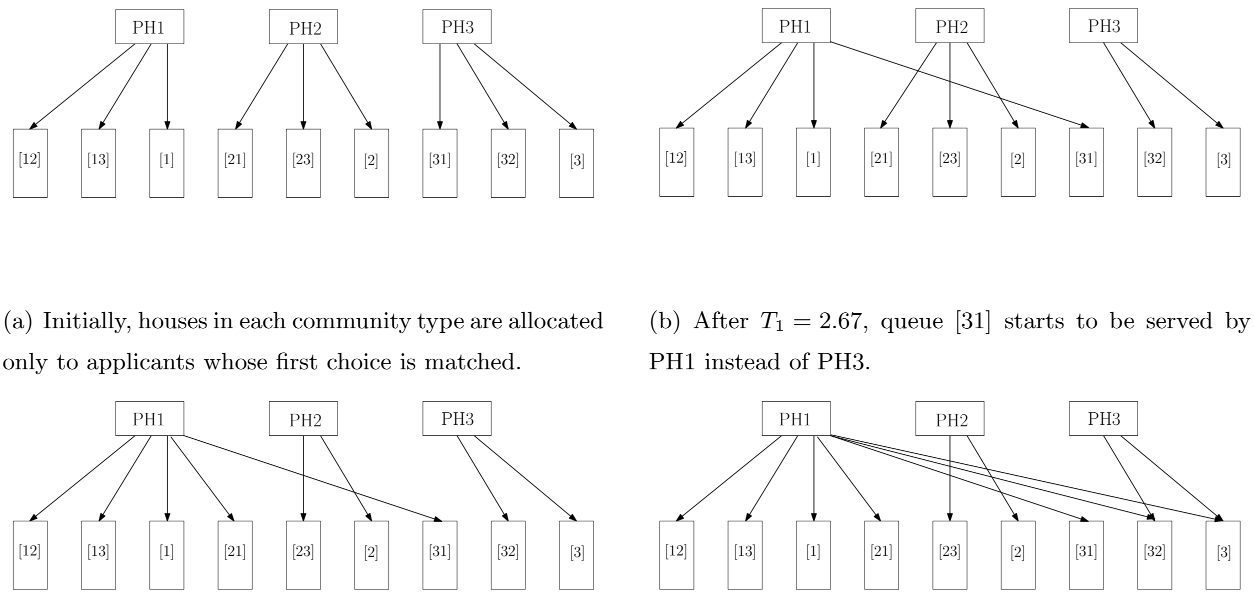
Figure 1 HOL Waiting Times for the nine queues predicted by Fluid Model (F) versus a Simulated Stochastic Process (S). Note that some queues, such as queues [32] and [3], have the same HOL waiting times curves. T_1, T_2, T_3 are switch times at which the routing components change. Other kinks on the curve are due to discontinuity of the arrival rate functions $\lambda(t)$ and $\mu(t)$.

4. Policy Evaluation

Our fluid model enables the housing authority to predict the properties of the system when housing allocations are made according to any M+W index in the following form,

$$\text{score}(j,i,\tau) = L(j,i) + g_i(\tau). \tag{2}$$

In practice, most housing authorities try to match the applicants' choice of community types to the available capacity, but an unintended consequence is that some community types might require a much longer waiting time than others. In practice, the queue length or waiting time is usually unobserved by the applicants. As a result, even applicants who are somewhat indifferent to the community type might have to wait much longer than the other applicants due to their



(a) Initially, houses in each community type are allocated only to applicants whose first choice is matched.

(b) After $T_1 = 2.67$, queue [31] starts to be served by PH1 instead of PH3.

(c) After $T_2 = 3.09$, queue [21] starts to be served by PH1 instead of PH2.

(d) After $T_3 = 19.15$, the routing components of PH1 and PH3 merge into a single routing component.

Figure 2 Routing configurations in each period

unintentionally selecting a housing type that corresponds to a longer queue. Such scenarios can be minimized if a housing authority actively tries to reduce the waiting-time disparity across different queues while aiming to match applicants' choices. To demonstrate how a HA can accomplish this, we assume that HACP tries to choose an M+W index that optimizes the following efficiency and fairness metrics,

$$\mathrm{Fa}(T) := -\sum_{i\in\mathcal{I}} \frac{\int_0^T \lambda_i(t)\,dt}{\int_0^T \sum_{i\in\mathcal{I}} \lambda_i(t)\,dt} \left(\frac{\int_0^T \sum_{j\in J} r_{ji}(t)dt}{\int_0^T \lambda_i(t)\,dt} - \frac{\int_0^T \sum_{j\in\mathcal{J}} \mu_j(t)\,dt}{\int_0^T \sum_{i\in\mathcal{I}} \lambda_i(t)\,dt} \right)^2. \tag{3}$$

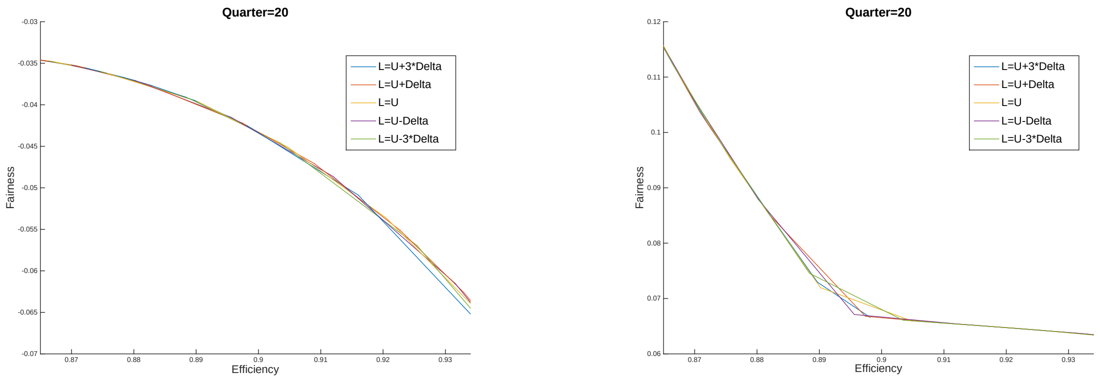
$$\mathrm{Ef}(T) := \frac{\int_0^T \sum_{i\in\mathcal{I}, j\in\mathcal{J}} U(j,i) r_{ji}(t)dt}{\int_0^T \sum_{j\in\mathcal{J}} \mu_j(t)\,dt}. \tag{4}$$

where $r(t) = r_{ji}(t)$ denote the service-rate matrix. The multi-objective optimization problem can be formulated as

$$\max \mathrm{Ef}(T) + \eta \mathrm{Fa}(T). \tag{5}$$

where η denotes the weight that the policy maker places on fairness against efficiency. We use the optimal index formula score* as the benchmark policy for our finite horizon problems for tractability. We then compare the performance of score* to several alternative formulas.

We can use Algorithm 4 in (Ding et al. 2018) to calculate the trajectories of queue-lengths or equivalently, head-of-line (HOL) waiting times, in a fluid model over a transient period. For example, as shown in Figure 3 (d), in Period 3, the corresponding residual network contains a cycle $PH1 \to [32] \to PH3 \to [3] \to PH1$; in Period 4, the cycle is even larger. We could construct alternative feasible routing rates by pushing a nonzero flow around these cycles, which would not change the matching utility because the utility around these cycles sums to zero. In other words, different routing rates lead to the same expected matching utility $\sum_{j\in\mathcal{J},i\in\mathcal{I}} r_{ji}U_{ji}$. Thus, in this example there is no need to perturb the matching score L. Indeed, our numerical experiments show that different perturbations of L lead to close performance with respect to the efficiency and fairness measurements (4) and (3).



(a) Fairness metric used: variance in getting offer

(b) Fairness metric used: lowest probability of getting offer

Figure 3 The Pareto frontiers when $L = U$, $U \pm \Delta$, $U \pm 3\Delta$ and $g_i(W(t)) = 1 - e^{-dW(t)}$. See (Ding et al. 2018) for definition of Δ.

We next fix the matching score to be $L = U$ and test the following parametric forms of the waiting score:

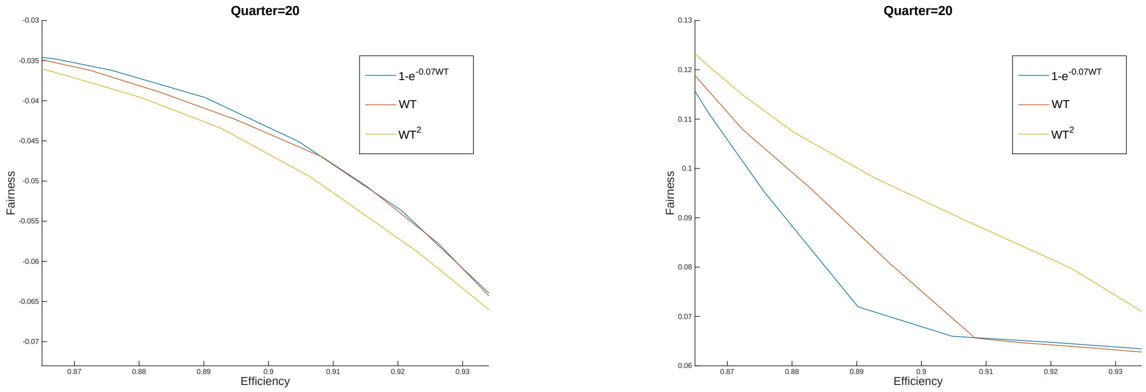
$$\begin{aligned} g_i(\tau) &= \tau \\ g_i(\tau) &= F_i(\tau) = 1 - \exp(-d\tau) \\ g_i(\tau) &= \tau^2 \end{aligned} \tag{6}$$

which represents linear, concave, and convex waiting scores. For a given weight η, we calculate the corresponding fluid process over a 20-quarter horizon. We measure efficiency using the average matching utility (4), but use two different fairness measurements. One is still the variance in likelihood of getting service given by (3), and the second is the minimum likelihood of getting service across all customers, which can be calculated by

$$\min_{i\in\mathcal{I}} \frac{\int_0^T \sum_{j\in\mathcal{J}} r_{ji}(t)dt}{\int_0^T \lambda_i(t)dt}. \tag{7}$$

By sampling different values of η, we obtain the Pareto frontiers corresponding to the three types of waiting scores and plot them on the efficiency-fairness spectrum in Figure 4, where 4 (a) uses the variance as fairness measure, and (b) uses the negative minimum likelihood as fairness measure. For the first fairness measure, we have a theoretical characterization of the optimal M+W index formula for the steady state, while the second fairness measure is much simpler and possibly more practical in real life. Figure 4 (a) shows that the status quo, $g_i(\mathrm{WT}) = F_i(W(t)) = 1 - \exp(-dW(t))$ outperforms the other two forms of waiting scores with respect to the first fairness measure; however, if the second fairness measure is used, then the convex waiting score $g_i(\mathrm{WT}) = F_i(W(t)) = 1 - \exp(-dW(t))$ dominates the other two. Intuitively, if one wishes to lower bound the likelihood of getting an offer, then one may consider using a convex waiting score to prioritize queues with the smallest likelihood.

In practice, the policy designer may wish to maximize the efficiency measure solely by taking the fairness measure as a constraint, for example, require customers in all queues to have a probability of at least 10% to get service. Then one may look up η at which the Pareto frontier intersects the line fairness $=$ 10% and obtain the corresponding M+W index.



(a) Fairness metric used: negative variance in getting offer

(b) Fairness metric used: lowest probability of getting offer

Figure 4 The Pareto frontiers when $L = U$ and $g_i(\mathrm{WT})$ takes the three forms in (6)

References

Ding, Yichuan, Thomas McCormick, Mahesh Nagarajan. 2018. A fluid model for an overloaded bipartite queueing system with heterogeneous matching utilities. *working paper* URL `http://blogs.ubc.ca/ycding/files/2018/08/PDF25891600-585444938.pdf`.

Geyer, Judy, Holger Sieg. 2013. Estimating a model of excess demand for public housing. *Quantitative Economics* **4**(3) 483–513.

Stockton, Halle. 2013. Pittsburgh, allegheny county experience critical shortage of public housing. *PublicSource* .