Group-level Fairness Maximization in Online Bipartite Matching*

Extended Abstract

Will Ma
Columbia University
New York, NY, USA
wm2428@gsb.columbia.edu

Pan Xu
New Jersey Institute of Technology
Newark, NJ, USA
pxu@njit.edu

Yifan Xu
Southeast University
Nanjing, CN
xyf@seu.edu.cn

ABSTRACT

We consider the allocation of limited resources to heterogeneous customers who arrive in an online fashion. We would like to allocate the resources "fairly", so that no group of customers is marginalized in terms of their overall service rate. We study whether this is possible to do so in an online fashion, and if so, what a good online allocation policy is.

We model this problem using online bipartite matching under stationary arrivals, a fundamental model in the literature typically studied under the objective of maximizing the total number of customers served. We instead study the objective of *maximizing the minimum service* rate across all groups, and propose two notions of fairness: long-run and short-run.

For these fairness objectives, we analyze how competitive online algorithms can be, in comparison to offline algorithms which know the sequence of demands in advance. For long-run fairness, we propose two online heuristics (Sampling and Pooling) which establish asymptotic optimality in different regimes (no specialized supplies, no rare demand types, or imbalanced supply/demand). By contrast, outside *all* of these regimes, we show that the competitive ratio of online algorithms is between 0.632 and 0.732. For short-run fairness, we show for complete bipartite graphs that the competitive ratio of online algorithms is between 0.863 and 0.942; we also derive a probabilistic rejection algorithm which is asymptotically optimal in the total demand.

Depending on the overall scarcity of resources, either our Sampling or Pooling heuristics could be desirable. The most difficult situation for online allocation occurs when the total supply is just enough to serve the total demand, in which case an organization could try to make allocations offline instead.

We simulate our algorithms on a public ride-hailing dataset, which both demonstrates the efficacy of our heuristics and validates our managerial insights.

KEYWORDS

Online Bipartite Matching; Fair Operations; Long-Run and Short-Run Fairness

ACM Reference Format:
Will Ma, Pan Xu, and Yifan Xu. 2022. Group-level Fairness Maximization in Online Bipartite Matching: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the online bipartite matching problem, nodes on one side of a bipartite graph are given in advance, while nodes on the other side arrive one-by-one. We refer to the two sets of nodes as *offline* and *online* agents, respectively. The edges incident to an online agent, which indicate the offline agents eligible to serve it, are revealed upon its arrival. An online matching algorithm must immediately serve each arriving agent using up to one eligible and unmatched offline agent; matches once made cannot be rearranged. The performance of an algorithm is determined by the total number of matches made, taking expectations as necessary if there is randomness in the arrivals or the algorithm. The *competitive ratio* (CR) measures the separation between the performance of online algorithms vs. that of a clairvoyant algorithm which knows all of the arrivals.

In this paper, we study online matching problems where performance is instead determined by the *fairness* in service provided to different groups of online agents. We assume that each online agent belongs to some protected groups, e.g. based on race or gender identity, which are observed upon arrival. To ensure that every group is adequately served, we evaluate performance by the *minimum* fraction of demand served over all the groups, defined in two different ways:

$$\text{Long-Run Fairness} = \min_{\text{groups } G} \frac{\mathbb{E}[\text{\# of agents in group } G \text{ served}]}{\mathbb{E}[\text{\# of arrivals in group } G]}; \qquad (1)$$

$$\text{Short-Run Fairness} = \mathbb{E}\left[\min_{\text{groups } G} \frac{\mathbb{E}[\text{\# of agents in group } G \text{ served}]}{\text{\# of arrivals in group } G}\right]. \qquad (2)$$

Motivation for Long-Run Fairness. The online matching time horizon represents a single day, and the algorithm is audited for fairness after a large number of days T have passed. In this case, the total number of group-j agents served over all the days will be statistically close to T times the numerator in (1), while the total number of group-j agents to arrive over all the days will be statistically close to T times the denominator. The audited performance is the minimum of this fraction over all groups j.

Motivation for Short-Run Fairness. The algorithm is audited for fairness based on the realized arrivals every single day. To avoid impossibility results[1], evaluation in the numerator of (2) is based on the *expected* service over any randomness in the algorithm.

*Pan Xu was partially supported by NSF CRII Award IIS-1948157. We thank the reviewers for their helpful comments. Here is the arXiv link to the full version: https://arxiv.org/abs/2011.13908.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online. © 2022 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

[1] Observe that any deterministic algorithm will yield a fairness of zero during peak hours when there are lots of groups each with a small arrival rate but the total rate is far larger than the serving capacity of offline agents.

Interpreted another way, when evaluating Short-Run Fairness, we are allowing for *fractional* allocations to be made on a given day. The overall performance (2) then takes the expectation of the daily audit scores over a large number of days.

Note that our objectives of Long-Run and Short-Run Fairness are percentages between 0 and 1. A guarantee on these percentages does not directly imply that all protected groups will enjoy an equitable level of service; however, these objectives naturally encourage algorithms to allocate the offline agents evenly across the online groups.

We acknowledge that our objectives for fairness at the group level do not address equity at the individual level [see 1, 2]; we make no considerations for the most "deserving" or "in need" agents within each group being served. Moreover, we are assuming that agents can be correctly labeled and there is no strategic behavior from individuals to obfuscate their groups. Nonetheless, we believe our objectives to be reasonable for large-scale online platforms, on which it has been found that under the current algorithms, agents in certain protected groups are significantly less likely to be served [3, 4].

We proceed with definitions (1)–(2) and answer the following questions:

(1) What is the fairness lost by imposing *non-rejection*, *i.e.,* that an online agent must be served (regardless of group) as long as there is an adjacent offline agent with remaining service capacity?
(2) In terms of maximizing fairness objectives (1) or (2), computing an optimal online policy may be hard, but can we derive simple, near-optimal online allocation heuristics?
(3) What is the competitive ratio, *i.e.,* the gap between the objective values (1) or (2) achievable by an online algorithm, vs. a clairvoyant offline algorithm which knows the arrival sequence in advance?

We believe Questions 1 and 3 to be particularly relevant for online platforms, addressing the design decision of whether incoming agents should be served whenever possible, and how much fairness the platform is losing by serving agents in an online instead of offline fashion. In this paper, we identify parameter regimes where a simple online heuristic achieves a competitive ratio approaching 100%, thereby also answering Question 2 in that it is a near-optimal online policy in these regimes.

2 MAIN CONTRIBUTIONS

In this paper, we assume that online agents arrive following independent Poisson processes with *known*, *homogeneous* rates. We see the assumption of rates being known as a modeling choice which puts us in the setting of online *stochastic* matching. On the other hand, our homogeneity assumption, that arrival rates do not change over time, does play a significant role in our results.

We now describe our results. For Long-Run fairness, we show that the competitive ratio of general online algorithms is between $1 - 1/e \approx 0.632$ (**Theorems 2, 4**) and $\sqrt{3} - 1 \approx 0.732$ (**Theorem 3**), while the competitive ratio of non-rejecting online algorithms is exactly $1/2$ (**Theorem 1**). Next, we establish that under specific parameter regimes, certain online heuristics achieve a competitive ratio approaching 1:

(1) When there are many copies of every offline agent, an online algorithm which *independently samples* an offline LP solution for each online agent achieves a competitive ratio approaching 1 (**Theorems 2, 4**);
(2) When all online agent types have a high arrival rate, an online algorithm which *pools and reserves* a set of offline agents to serve each online agent type achieves a competitive ratio approaching 1 (**Theorem 5**);
(3) When a demand saturation parameter s^* approaches 0 or ∞, the LP sampling algorithm achieves a competitive ratio approaching 1, assuming that every protected group G is *homogeneous*, *i.e.,* consists of a single online agent type (**Theorem 2**).

For Short-Run Fairness, we assume there to be b copies of a *single* offline agent, which can be interpreted as one divisible resource. We show that the non-rejecting First-Come-First-Serve algorithm achieves a competitive ratio of 0.863, when the total arrival rate Λ of online types is at most 1. On the other hand, we derive a *probabilistic rejection* algorithm which is asymptotically optimal as $\Lambda \to \infty$, with b allowed to depend arbitrarily on Λ. We note that this algorithm performs rejections using randomness that is *dependent* across agents, making it different from the independent sampling algorithm mentioned earlier. Finally, we show that the competitive ratio of online algorithms is upper-bounded by 0.942, even when $b = 1$. All details regarding the Short-Run Fairness are deferred to the full version.

3 EXPERIMENTS ON RIDE-HAILING DATASETS

Using a ride-hailing dataset collected from the city of Chicago[2], we test our heuristics against existing algorithms in the Online Bipartite Matching literature, in some cases adapting them for our Long-Run Fairness objective. We consider both the general case where protected groups (of riders, based on origin and destination of trip) can consist of heterogeneous types, and the special case where protected groups consist of a single type (*i.e.,* a group is defined by a single origin and destination pair). Our findings are summarized below.

First, in the case of homogeneous groups, our sampling heuristic always achieves higher Long-Run Fairness than the existing Online Matching algorithms, over a range of choices on how to scale the demand saturation. Moreover, the general performance of all the algorithms is *exactly consistent* with the managerial insights from our theory—the most difficult situation for achieving fairness in an online fashion arises when the total supply and demand in the system is balanced. On the other hand, all online algorithms perform better relative to the optimal offline allocation when the supply-demand imbalance increases (in either direction).

Second, in the case of heterogeneous groups, online matching using our sampling heuristic is effective if the minimum supply capacity is large, while online matching using our pooling/pre-reserving heuristic is effective if the minimum demand rate is large. These observations from data are also consistent with our algorithmic guarantees and managerial insights.

[2] https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p

REFERENCES

[1] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.

[2] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[3] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* 9, 2 (2017), 1–22.

[4] Jorge Mejia and Chris Parker. 2020. When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* (2020).