On Fairness and Stability in Two-Sided Matchings

Gili Karni
Weizmann Institute of Science, Rehovot, Israel

Guy N. Rothblum
Weizmann Institute of Science, Rehovot, Israel

Gal Yona
Weizmann Institute of Science, Rehovot, Israel

Abstract

There are growing concerns that algorithms, which increasingly make or influence important decisions pertaining to individuals, might produce outcomes that discriminate against protected groups. We study such fairness concerns in the context of a two-sided market, where there are two sets of agents, and each agent has preferences over the other set. The goal is producing a matching between the sets. Throughout this work, we use the example of matching medical residents (who we call "doctors") to hospitals. This setting has been the focus of a rich body of work. The seminal work of Gale and Shapley formulated a *stability* desideratum, and showed that a stable matching always exists and can be found in polynomial time.

With fairness concerns in mind, it is natural to ask: might a stable matching be discriminatory towards some of the doctors? How can we obtain a *fair* matching? The question is interesting both when hospital preferences might be discriminatory, and also when each hospital's preferences are fair.

We study this question through the lens of metric-based fairness notions (Dwork *et al.* [ITCS 2012] and Kim *et al.* [ITCS 2020]). We formulate appropriate definitions of fairness and stability in the presence of a similarity metric, and ask: does a fair and stable matching always exist? Can such a matching be found in polynomial time? Can classical Gale-Shapley algorithms find such a matching? Our contributions are as follows:

- **Composition failures for classical algorithms.** We show that composing the Gale-Shapley algorithm with fair hospital preferences can produce blatantly unfair outcomes.
- **New algorithms for finding fair and stable matchings.** Our main technical contributions are efficient new algorithms for finding fair and stable matchings when: (*i*) the hospitals' preferences are fair, and (*ii*) the fairness metric satisfies a strong "proto-metric" condition: the distance between every two doctors is either zero or one. In particular, these algorithms also show that, in this setting, fairness and stability are compatible.
- **Barriers for finding fair and stable matchings in the general case.** We show that if the hospital preferences can be unfair, or if the metric fails to satisfy the proto-metric condition, then no algorithm in a natural class can find a fair and stable matching. The natural class includes the classical Gale-Shapley algorithms and our new algorithms.

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains

Keywords and phrases algorithmic fairness

Digital Object Identifier 10.4230/LIPIcs.ITCS.2022.92

Related Version *Full Version*: `http://arxiv.org/abs/2111.10885`

Funding *Gili Karni*: Research supported by the Israel Science Foundation (grant number 5219/17), and by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.
Guy N. Rothblum: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), from the Israel Science Foundation (grant number 5219/17), and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

Gal Yona: Supported by the European Research Council (ERC) (grant agreement No. 819702), by the Israel Science Foundation (grant number 5219/17), by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, by the Israeli Council for Higher Education via the Weizmann Data Science Research Center, by a Google PhD fellowship, and by a grant from the Estate of Tully and Michele Plesser.

Acknowledgements We thank Shahar Dobzinski and Moni Naor for helpful conversations and suggestions about this work and its presentation. We also thank the anonymous ITCS 2022 reviewers for their feedback.

1 Introduction

Algorithms are increasingly influencing or replacing human decision-makers in several sensitive domains, including the criminal justice system, online advertising, and medical risk prediction. Along with the benefits of automated decision-making, there is also a potential risk of discrimination towards groups of individuals, which might be illegal or unethical [22, 3]. Examples for unintended but harmful behavior have been shown to happen in algorithms that allocate resources in domains such as online advertising [1, 29], healthcare systems [21, 23], and more.

Many resource allocation problems, such as online advertising and assigning hospitals to medical students for residency, can be viewed as *two-sided markets*. In this setting, there are two sets of agents that both have preferences over the other set. We seek a matching: a symmetric allocation in which every member from the first set is matched to a member of the other set (ads to users, drivers to customers, hospitals to students). A well-studied desideratum for such two-sided matchings is *stability* [12]. A matching is stable only if it leaves no pair of agents on opposite sides of the market who are not matched to each other, but would both prefer to be. In their seminal work, Gale and Shapley [12] proved that in two-sided markets, a stable matching always exists and can be found efficiently using a simple procedure known as the *Gale-Shapley* algorithm. Stability represents the incentive to participate in the matching, i.e., no unmatched pair can *both* improve their situation by being matched to each other. This problem was introduced by [12] and has been studied broadly, see e.g. the books by Roth and Sotomayor [26], by Knuth [20], and by Gusfield and Irving [14]. In particular, Roth [24, 25] studied the real-world setting of matching medical residents to hospitals. We use this as a running example throughout our work.

There are many instances, however, where we may seek to look beyond utility-based desiderata such as stability. Suppose, for example, in the problem of assigning hospitals for residency, that given two equally-qualified residents, the preferences of some hospitals are discriminatory, e.g., they display a preference towards residents who do not live in certain neighborhoods, or residents without children. In this case, a stable matching may be undesirable. Our goal is preventing discrimination in the two-sided matching setting, where we match individuals to resources such as hospitals, ads, or schools. The emerging literature about algorithmic fairness typically focuses on one-sided allocation problems (e.g., supervised learning). The two-sided setting is unique in that we have two sets of agents, both of which have preferences over the other side. Studying fairness in the context of two-sided markets is well-motivated since many of the resource allocation problems captured by two-sided markets have high stakes for the individuals involved.

1.1 This Work: Fair and Stable Matchings

We would like the matching to be both fair and stable. This requires appropriate definitions of fairness and stability. In this work, we embark on a study of this question; we set out to present such definitions and study when and how fairness and stability can be compatible.

A fundamental prerequisite is a notion of what makes a given allocation fair. We build on the approach of *individual fairness* (IF) [7], which assumes the existence of a task-specific similarity metric that measures how "similar" two individuals are (say, how similarly qualified they are). We would like to ensure that the eventual allocation satisfies *preference-informed individual fairness* (PIIF) [17], which roughly means that there is no envy between similar individuals (i.e., a doctor i_1 will not prefer the outcome that a similarly qualified doctor i_2 receives). In the case where similar doctors have the same preferences, a fair deterministic solution does not exist. Thus, we focus on finding a fair *distribution* over matchings.

In this work, we begin to chart the landscape of *fair and stable* matchings in the context of two-sided markets. Specifically, our contributions are:

- **Fair hospitals might arrive at unfair allocations**. Our exploration builds on an important but counter-intuitive result. We show that even when the hospitals' preferences satisfy a very strong notion of fairness (strong indifference between equally-qualified candidates), running the classic Gale-Shapley algorithm is *not* guaranteed to result in a fair allocation. This can be seen as a failure of *composition*: even when the inputs to an algorithm are "fair", the result may not be. Composition has previously been highlighted as a challenge in the context of designing fair algorithms [8]. This stands in stark contrast to e.g. the landscape around privacy (where *differential privacy* enjoys graceful composition properties, facilitating the design of complex private algorithms using private "building blocks").
- **New algorithms assuming fair preferences and "simple" metrics**: Our main technical contribution is a strong positive result establishing that fairness and stability are compatible, when (i) the preferences of the hospitals are fair, and (ii) we place strong assumptions on the structure of the similarity metric defining fairness. Specifically, our results hold for a class of similarity metrics that we refer to as "proto-metrics", in which the distances between every pair of individuals must be either zero or one. Importantly, we show that not only do fair and stable solutions exist, but they can be found efficiently: we provide new algorithms, inspired by the Gale-Shapley algorithm, which obtain fair and stable solutions.
- **Barriers for compatibility in the general case**. A natural question is whether the assumptions that the hospital preferences must be fair and that the metric must be "simple" are necessary. We provide a partial answer by demonstrating a rich class of natural algorithms (extending our algorithms and the original Gale-Shapley algorithm), and proving that no algorithm in this class can guarantee a fair and stable solution if either assumption is removed.
- **New notions of stability**. The above barrier results require formalizing new notions of stability for this more general setting. We aim to formalize notions of stability that provide a reasonable guarantee for the hospitals without being trivially incompatible with fairness. For example, if the hospital that is preferred by all the doctors has discriminatory preferences, it might be the case that every stable matching is blatantly unfair. Similar situations can arise even when the hospital preferences are fair, if the metric is not a proto-metric. Thus we aim for a relaxed goal, which we find suitable for this setting: obtaining a *fair matching* where no hospital prefers any *fair alternative* to the allocation it received in the matching. Formalizing this intuition presents several subtleties. Thus,

we present two definitions, one for the case of unfair preferences and a proto-metric, and a weaker definition for the case of general metrics (we use this weaker definition in our negative results).

1.2 Related Work

Our work bridges the established literature on matchings, originating in the seminal work of Gale and Shapley [12], and the emerging literature on algorithmic fairness, which seeks to formalize and mitigate discrimination in algorithmic decision-making systems [2]. The latter has focused on formalizing and studying different notions of fairness, and understanding the tensions between them and possible "accuracy" based desiderata [6, 19, 15]. One popular approach to quantifying unfairness are group-based definitions, in which the objective is equalizing some statistic across a fixed collection of "protected" groups. In this work, we build on a different approach, using the notion of preference-informed individual fairness [17], which combines the individual-based fairness notion proposed in [7] with the notion of envy freeness from game theory [10, 30]. Recent work [8] has considered the question of composition in the context of fairness, showing that fairness may fail to compose in a variety of settings. Our results complement their findings by demonstrating that such failures may also occur for a natural and widely popular algorithm.

Fairness has also been studied previously in the context of two-sided matchings. [14] introduced the *equitable, stable marriage problem*, where the objective is to minimize (across stable matchings) the difference between the sum of rankings of each set over their matches. This implies avoiding unequal degrees of happiness among the two sides, and [13] showed an algorithm that finds such a solution. This is motivated by the fact that stable matchings are in general not unique (in fact, their number could even be exponential in the size of the sets [20, 28]), and that the algorithm introduced by [12] finds the optimal stable matching for one of the sets and the worst stable matching for the other. Our approach is inherently different in that we seek a matching that is fair within one of the sets rather than across sets. [11] studied the problem of finding an allocation that satisfies *envy-freeness up to one good* (EF1) for both sets simultaneously. This is a relaxation of the classic notion of envy-freeness, which requires that any pairwise envy can be eliminated by removing a single item from the envied agent's allocation. This formulation is similar to ours in that it tries to guarantee fairness for each set separately. However, our work is conceptually different in that we consider metric-based fairness requirements as well as our focus on the compatibility between fairness and stability, and technically different because, in the one-to-one setting that we study, the EF1 objective becomes vacuous. Finally, [27] study the problem of matching drivers to passengers in ride-hailing platforms, trying to achieve equal income for the drivers. We focus on a different fairness notion that is appropriate when the preferences of the individuals are potentially diverse.

2 Defining Fairness and Stability

We focus on an asymmetric setting, such as assigning hospitals to medical students for residency. Usually, in this problem, each hospital can be assigned to multiple residents. For simplicity, we focus on the setting where each hospital is assigned to a single resident. We refer to the medical students as doctors. We want to guarantee fairness for the doctors and stability, to be defined, for the hospitals. Since the resources are limited and sometimes multiple individuals want a single resource, a deterministic fair solution does not always exist. Thus, we focus on finding probabilistic solutions where the allocation is a distribution over matchings.

2.1 Fairness Requirements

To enforce fairness, we assume that we are given an unbiased similarity metric for the doctors. We wish to ensure a metric-based fairness guarantee such as *individual fairness* (IF) [7], that is, similar individuals should have a similar outcome. For instance, the metric can represent differences in the GPA of medical students. Then, students with a GPA of 5.0 should be assigned to a prestigious hospital with the same probability, and students with a GPA of 4.0 can be assigned to that prestigious hospital with a lower probability. However, sometimes similar individuals have different preferences, e.g., students with the same GPA can prefer different hospitals for reasons such as geographic location or specialization in a particular field. In that case, we want to allow these similar individuals to have different outcomes. Thus, we find that *preference-informed individual fairness* (PIIF) [17] is a more appropriate fairness notion. PIIF is a relaxation of two fairness requirements: (1) individual fairness; and (2) envy freeness. Envy freeness (EF) requires that no individual prefers the outcome of another individual over their own outcome. In PIIF, we allow deviations from EF, i.e., we allow for an individual i to envy the outcome of another individual j, conditioned on the fact that j's outcome can be changed by no more than the distance between i and j to an outcome that i will not envy. We allow deviations from IF, so long as they are in line with individuals' preferences. See Definitions 7, 8, and 9 for formal definitions of IF, EF, and PIIF.

2.1.1 Our Focus: Proto-Metrics

In many of our results, we focus on the special case where the metric is a *proto-metric*: the distances between all the doctors are either 0 or 1. In this setting, the doctors are divided into clusters of similar doctors. PIIF means that we require envy-freeness between the doctors in each cluster (see Corollary 10). However, there are no constraints on the allocations of doctors from different clusters. We focus on this setting throughout this extended abstract, *except* in Section 5.

To compare the preferences of the doctors within each cluster, we use *stochastic domination*. We say one distribution stochastically dominates another if, for every outcome, the probability of getting this outcome or a better one is lower-bounded by the corresponding probability in the other distribution. See Definition 6 for a formal definition of stochastic domination. In the presence of a proto-metric, a PIIF allocation implies that each doctor's allocation stochastically dominates the allocation of any other doctor in its cluster.

We find that there are natural settings where the restriction to proto-metrics is reasonable. For example, consider a setting where hospitals are only allowed to distinguish between medical students based on a specialization in their medical studies, e.g., neurology, cardiology, etc.[1] Different hospitals may prefer different specializations. A proto-metric can partition the candidates based on their specialization (one could also add merit-based sub-categories within each specialization).

[1] For example, in the current Israeli system, hospitals are not allowed to express preferences over the medical students [5]. Allowing hospitals to have fair preferences, even in a limited way, could improve outcomes for the hospitals.

2.2 Fair Hospital Preferences

We distinguish between two scenarios: (1) Hospitals have fair preferences over the doctors. For instance, we allow a hospital to prefer some doctors over others because they have a higher GPA or good grades in a particular topic but not because they do not have children or belong to a specific ethnic group. The fairness requirement for the hospitals' preferences can be formalized in different ways (see below). For now, we emphasize that, even if we impose strict fairness requirements on the individual hospitals' preferences, our results show that obtaining a fair and stable solution can be far from trivial. (2) Hospital preferences might be discriminatory. This case is interesting since decision-makers can be discriminatory, e.g., because of biased data or prejudice. Finding a fair and stable solution in the presence of unfair hospital preferences is even more challenging (in particular, it is not clear how to define stability), see Section 5.

Requiring fair hospital preferences is an assumption or restriction on the input to the algorithm that attempts to find a fair and stable allocation (the algorithm's input is the preferences of the doctors and the hospitals). We find the restriction to be natural and well motivated: if our goal is finding a matching that is fair to the doctors, it makes sense to ask the hospitals to indicate fair preferences. As remarked above, even under strong fairness restrictions on the hospitals' preferences, finding a fair and stable matching is challenging.

2.2.1 Our Focus: Strictly Fair Hospital Preferences

Focusing on the proto-metric setting (see above), we formalize a strong notion of *strictly individually fair* hospital preferences: each hospital can have arbitrary (deterministic) preferences over the clusters, but must be completely indifferent between every two doctors that are in the same cluster.

To reason about the (metric-based) fairness of a hospital's preferences, we view them as probabilistic, i.e., a distribution over ordinal preferences. Strictly IF preferences induce such a distribution, where the "external" ordering of the clusters is deterministic, and the "internal" ordering within each cluster is a uniformly random permutation of the doctors (the random internal ordering captures indifference between doctors in the same cluster).[2]

2.2.2 Beyond Strictly IF Preferences

Strict IF is a strong restriction or assumption on the hospitals' preferences. This makes our negative results (the failure of the classic Gale Shapley algorithms) stronger: the algorithms fail even under the strong restriction on hospital preferences. For positive results, showing an algorithm that works given a more relaxed notion of fair preferences (even in the proto-metric setting) is an interesting question for future work.

More generally, we could hope to design algorithms that work with fair hospital preferences under general metrics, or with completely *unfair* hospital preferences. This raises subtle difficulties in the definition of stability and encounters natural barriers, see Section 5.

[2] We note that while strict IF could also be captured with deterministic preferences that allow ties, as was considered in [16, 18], reasoning about preferences as distributions over ordinal preference lists allows extensions to other notions of fair preferences, as well general metrics.

2.3 Stability Requirements

The uniform distribution over all outcomes is always trivially fair. However, in addition to fairness for the doctors, we want some guarantee for the hospitals. Stability is one such guarantee, which has been studied extensively in the classical setting, with deterministic allocations, and without fairness constraints. The classical stability guarantee ensures that there are no "blocking pairs". That is, there are no pairs of a hospital and a doctor that prefer each other over their match and are not matched. We extend this notion to the setting where there are fairness constraints, and the allocations are probabilistic.

2.3.1 Our Focus: Stability Under Proto-Metrics, Strictly IF Hospital Preferences

For the case of a proto-metric and strict IF hospital preferences, there is a natural extension to the classical notion of stability. We say that a (probabilistic) allocation is unstable if there is a matching with non-zero probability under the probabilistic allocation, where there exists a blocking pair (d, h) that (strongly) prefer each other over their match. We remark that, since we assume the hospital preferences are strictly IF, this can only happen if the doctor d and the doctor who is matched to the hospital h (denote them d'), are in different clusters (otherwise h, whose preferences are strictly IF, will not prefer d to d').

For instance, suppose there two clusters $i = \{i_1, i_2\}$ and $j = \{j_1, j_2\}$, such that hospital h prefers cluster i over cluster j, and doctor i_1 prefers hospital h over any other hospital. Assume a probabilistic allocation where, for some matching in the support, hospital h is matched to j_1. The pair h and i_1 form a blocking pair, since they strongly prefer each other over their match. On the other hand, even if hospital h is matched to doctor i_2 in every matching in the support, hospital h and doctor i_1 do not form a blocking pair, since hospital h does not have a strong preference between doctors i_2 and i_1 (since they are in the same cluster).

2.4 Approximate Stability

For a probabilistic allocation, we can relax the stability requirement by allowing a blocking pair to occur with a small probability. The probability is over the choice of a matching drawn from the probabilistic allocation. We say that an allocation is τ-approximately stable if the probability that *no* blocking pair occurs is at least $(1 - \tau)$.

2.4.1 Beyond Proto-Metrics and Fair Preferences

Defining an appropriate notion of stability under unfair hospital preferences or general metrics is considerably more challenging. We provide a definition for unfair hospital preferences in the presence of a proto-metric. We also formalize a minimal weak stability requirement for general metrics (we use this requirement to show negative results). See Sections 5.1 and 5.3.

3 Fair Preferences Do Not Guarantee a Fair Allocation

Focusing on the setting of a proto-metric and strictly IF hospital preferences, a natural way to achieve stability and fairness is to use the probabilistic form of strict IF preferences (see Section 2.2), sample the hospitals' preferences and run the Gale-Shapley algorithm over these sampled preferences. The probabilistic allocation is the random variable defined by

this procedure. In the full version of the paper, we show that while this produces a stable probabilistic allocation, it can lead to unfair outcomes, even when the hospital preferences are themselves strictly IF.

To explain this negative result, we first present the Gale-Shapley algorithm. The algorithm is not symmetric: one of the sets is the proposing set, and the other is the accepting set. Here we present the variant where the doctors are the proposing set: At the initialization, no doctor is matched. The algorithm terminates when each doctor is matched to a hospital. Until then, at each round, an unmatched doctor d is chosen arbitrarily. This doctor d proposes to a hospital h, where d chooses h as the most preferred hospital that did not reject it yet. Then, hospital h has to decide whether to accept doctor d or reject it. If hospital h is unmatched too, it will always accept. If hospital h is already matched to a doctor d', it will accept only if it prefers doctor d over doctor d'. Otherwise, it will reject. If hospital h accepts doctor d, it rejects doctor d'.

▶ **Theorem 1** (Running Gale-Shapley over fair preferences Gale-Shapley over fair preferences does not guarantee fairness (informal))**.** *The algorithm that generates a probabilistic allocation by sampling hospital preferences from a strict individually fair distribution and running the Gale-Shapley algorithm over the sampled preferences is not fair.*

Proof sketch. Consider the example of three doctors i_1, i_2, j, three hospitals A, B, C, and the proto-metric d, where $d(i_1, i_2) = 0$ and $d(i_1, j) = d(i_2, j) = 1$. The doctor preferences are

$$A \succ_{i_1} B \succ_{i_1} C, \quad A \succ_{i_2} C \succ_{i_2} B, \quad C \succ_j A \succ_j B.$$

The hospitals have strictly individually fair preferences: (note that i_1 and i_2 are interchangeable)

$$\begin{cases} j \succ_A i_1 \succ_A i_2, & w.p.\ 1/2 \\ j \succ_A i_2 \succ_A i_1, & w.p.\ 1/2 \end{cases}, \quad \begin{cases} j \succ_B i_1 \succ_B i_2, & w.p.\ 1/2 \\ j \succ_B i_2 \succ_B i_1, & w.p.\ 1/2 \end{cases}, \quad \begin{cases} i_1 \succ_C i_2 \succ_C j, & w.p.\ 1/2 \\ i_2 \succ_C i_1 \succ_C j, & w.p.\ 1/2 \end{cases}.$$

Running the algorithm described above that samples the preferences and uses the doctor-propose Gale-Shapley algorithm results in the allocation

$$\begin{cases} (i_1, B), (i_2, A), (j, C), & w.p.\ 1/2, \\ (i_1, B), (i_2, C), (j, A), & w.p.\ 1/2. \end{cases}$$

Hospital A is the most preferred hospital by doctor i_1, but doctor i_2 (who is similar to doctor i_1) is matched to hospital A with higher probability. Thus, this allocation is unfair. ◀

For a more detailed discussion, see the full version of this paper. In the full version of the paper, we show a negative example for the hospital-propose Gale-Shapley variant. In these examples, although all the participants acted fairly, the outcome was unfair; this joins existing work on fairness failures under composition [8, 9].

3.1 Digest: Towards Fairness

In the counter-example presented above, when in the sampled preferences hospital A prefers doctor i_2 over doctor i_1, the algorithm matches hospital A to doctor i_2, without making sure that in the corresponding case, where hospital A prefers doctor i_1 over doctor i_2, it would match hospital A to doctor i_1. Intuitively, we would like the algorithm to make this decision

simultaneously, i.e., to either match both doctors i_1 and i_2 to hospital A with a certain probability or not to match neither of them. More generally, when matching a doctor d to a hospital h, we must ensure that all the doctors in d's cluster have the same opportunity to be matched to this hospital.

4 Positive Results: Algorithms for Fair and Stable Allocations

In the full version of the paper, we present generalizations of the Gale-Shapley algorithm that achieve both fairness and stability, up to a small error.

▶ **Theorem 2** (Compatibility of fairness and stability (informal)). *There exists an efficient algorithm that, given a proto-metric, strictly IF hospital preferences, arbitrary doctor preferences, and an approximation parameter $\tau \in (0,1)$, always finds a τ-approximately fair and τ-approximately stable allocation. The algorithm's running time is polynomial in $(1/\tau)$.*

The allocation returned by the hospital-first variant of the algorithm is (perfectly) PIIF and τ-approximately stable. The allocation returned by the doctor-first variant of the algorithm is τ-approximately PIIF and τ-approximately stable. See the definition of τ-approximate stability in Section 2.3. τ-approximate fairness means that for every pair i_1, i_2 of doctors in the same cluster, i_2's allocation is τ-close (in statistical distance) to an allocation that i_1 doesn't envy. See Definitions 16 and 11 for formal definitions of approximate stability and approximate fairness.

In these algorithms, we allow the parties to propose probability mass to each other. In the variant where the hospitals propose, they propose to clusters instead of individuals, so that all doctors in the cluster have an equal opportunity to accept. In the variant where the doctors propose, the hospitals accept an allocation that will not cause envy within clusters, e.g., in the counter example of Theorem 1, hospital A can either accept both doctors i_1 and i_2 or reject both of them. If the distances between all the doctors are 1, these algorithms are identical to the Gale-Shapley algorithm. The formal description of the algorithms and the full proofs are in the full version of the paper.

4.1 Overview: Fair Propose-and-Reject – Doctors-First

At the initialization of the algorithm, each doctor has a free probability mass of 1. At each round, every doctor proposes its free probability mass to the most preferred hospital that did not reject it yet. For every hospital h, after the doctors' proposals, the probability mass proposed by each doctor is composed of the probability mass the doctor proposed in the current round, plus the probability to be matched to this doctor in the previous round. Given this proposed probability mass, the hospital chooses its allocation for the current round in a way that will not lead to unfairness.

In particular, each hospital h uses the *rising tide algorithm* (described in the full version of the paper) for choosing its allocation in each round. At the initialization, hospital h has unallocated probability mass 1. The algorithm goes from the most preferred cluster to the least preferred cluster by hospital h. When there is no unallocated probability mass or no proposed probability mass, the algorithm terminates. For each cluster C, while there is proposed probability mass from cluster C: Each doctor d in cluster C is allocated with the minimum between: (1) the proposed probability mass from doctor d; (2) the unallocated probability mass divided by the number of doctors in the cluster with non-zero proposed probability mass. Every allocated probability mass is removed from the proposed probability mass and from the unallocated probability mass.

After running the algorithm described above, in the allocation of the current round, hospital h is allocated to the most preferred clusters possible given the proposed probability mass. For every two doctors in the same cluster, one can be allocated with less probability mass than the other only when having less proposed probability mass.

Any probability mass not used for the allocation is rejected and becomes free again for the next round. The algorithm terminates when there is no more free probability mass, i.e., there is a full allocation, or the free probability mass is very small.

This algorithm is executed *locally*, in the sense that for each hospital, the only necessary knowledge is its own preferences, its previous allocation, and the proposals it received. For each doctor, the only necessary knowledge is its own preferences, its free probability mass, and the hospitals' response. No party needs to be aware of the status of the other parties.

4.2 Overview: Fair Propose-and-Reject – Hospitals-First

This algorithm is similar to Algorithm 4, except that the hospitals propose their free probability mass to the clusters. At each round, every hospital proposes to the most preferred cluster that did not reject it. For each cluster, the proposed probability mass from each hospital is composed of the probability mass the hospital proposed in the current round, plus the probability to be matched to this hospital in the previous round.

Given this proposed probability mass, each cluster chooses an envy-free allocation for the current round. To achieve envy freeness, we use the *probabilistic serial procedure* due to Bogomolnaia and Moulin [4]. Any probability mass not used for the allocation is rejected and becomes free again for the next round.

The algorithm terminates when there is no free probability mass, i.e., there is a full allocation, or the free probability mass is very small.

This algorithm is also executed locally: no party has to consider anything but its own interests (though there is some coordination in each round between the doctors in the same cluster, in allocating the mass proposed to that cluster).

5 Barriers for Unfair Preferences and General Metrics

In the full version of the paper, we show that if we relax the strong requirement on the hospitals' preferences by allowing either unfair preferences or a general metric, no algorithm that is "similar in spirit" to the algorithms outlined above (and to the classical Gale-Shapley algorithm) can guarantee both fairness and stability. Towards this, we formalize a class of *local-proposing algorithms*. Here, we focus on the case where the doctors propose, although we also describe a class of hospital-proposing algorithms. The doctor-proposing class includes all algorithms consisting of sequential rounds of proposals. In each round, each doctor proposes its unallocated probability mass to a hospital. Then, each hospital has to use the probability mass proposed to it to choose an allocation. We assume that doctors choose the hospital according to their preferences. We also assume that once a doctor proposes some probability mass to a hospital, if the hospital rejects it, the doctor will never propose it to this hospital again. This class, and the corresponding hospital-proposing class, generalizes the algorithms outlined above (as well as the classical Gale-Shapley algorithms).

To provide this negative result, we need definitions of stability that extend beyond the setting of strictly IF hospital preferences. Remaining in the proto-metric setting, we define stability under unfair hospital preferences in Section 5.1. The negative results for local algorithms under unfair preferences are in Section 5.2. Moving to general metrics, we formalize

a minimal stability requirement in Section 5.3.1 (a weaker stability requirement makes our negative results stronger). We discuss our negative result for fair hospital preferences under general metrics in Section 5.4.

5.1 Stability under Unfair Preferences, Proto-Metric

If the hospitals can have unfair preferences, then fairness and stability might be trivially incompatible. For example, a prestigious hospital h can express discriminatory preferences towards members of a group T. Suppose h is the most-preferred hospital of all doctors: a stable allocation must always match h to a member of T, but this is blatantly unfair!

We want to provide a utility guarantee to the (unfair) hospitals, but the example above demonstrates that if we allow unfair hospitals to make unfair deviations from their allocation, then fairness and stability might be trivially incompatible. Since we insist on fairness for the doctors, we find that it is natural to relax stability by requiring that there are no *fair* deviations that the hospitals would prefer. Formally, we modify the classical notion of a blocking pair, by only allowing a hospital to form pairs with doctors that are *outside* the cluster of the doctor to whom it is matched. This restricts the alternative "offers" that a hospital with unfair preferences can make, and ensures that they do not violate the fairness constraints.

▶ **Definition 3** (Stability under proto-metrics (informal)). *A probabilistic matching is unstable if there exists, with non-zero probability over the matching, a pair of a doctor d and a hospital h that prefer each other to their respective partners, as long as the partner of the hospital h is at distance 1 from the doctor d. The preferences of the hospital can be probabilistic. Thus, the comparison between the hospital's allocations is in terms of stochastic domination.*

See Section 2.1, for a definition of stochastic domination.

See Definition 15 for a formal definition of stability. We note that if the hospitals' preferences are strictly IF, no hospital prefers one doctor over the other if they are in the same cluster. Thus, Definition 3 is equivalent to the definition described in Section 2.3.

5.2 Barriers for Unfair Preferences, Proto-Metric

We show that no local-proposing algorithm can find a fair and stable matching when the hospital preferences can be unfair (even in the proto-metric setting).

▶ **Theorem 4** (Failure of local algorithms for unfair hospital preferences (informal)). *There does not exist a local algorithm that, when the metric is a proto-metric, but the hospital preferences might be unfair, always finds an allocation that is both PIIF and stable (see Definition 3).*

Proof sketch. Suppose there are four doctors i, j, k, l, where the first and last pairs are at distance 0 (i.e., (i, j) and (k, l)), and the rest are at distance 1. Suppose there exists a hospital A, whose preferences are $j \succ_A k \succ_A i \succ_A l$. If hospital A is the most preferred hospital by all the doctors, then the fair allocation that is most preferred by hospital A is to be assigned to doctors i and j uniformly (the other fair possibilities are to be assigned to doctors k and l uniformly, or to some convex combination of these two allocations). Thus, since it is also doctors i and j's most preferred fair allocation (they both rank A first), it is the only stable one (for any other fair allocation, hospital A and doctor j form a blocking pair). However, if hospital A is the most preferred hospital by doctors k and i, but doctors j and l are matched to hospitals they prefer over hospital A. Then, the fair allocation that is most preferred by hospital A is to be assigned to doctor k with probability 1 (assuming

hospital A cannot be assigned to doctor j). Similarly, since this is also doctor k's most preferred allocation (doctor k rank hospital A first), it is the only stable one, otherwise hospital A and doctor k form a blocking pair.

If only the doctors i and k propose to hospital A probability mass 1 in the first round, there is no allocation that hospital A can choose over this probability mass that will always lead to a fair and stable matching, i.e., hospital A does not know whom to accept and whom to reject. If doctors j and l will propose to hospital A in a later round, to have a stable and fair solution, hospital A must accept at least probability mass $1/2$ from doctor i. However, if no other doctor will propose to hospital A in a later round, to have a stable and fair solution, hospital A must accept probability mass 1 from k. Since hospital A cannot distinguish these two cases in the first round, the algorithm will return an unfair or unstable output in at least one of them. ◀

Unfair preferences allow us to create a situation where hospital A actually wants to accept probability mass from doctor j, but under the fairness requirement, in order to accept probability mass from doctor j, hospital A must accept some probability mass of doctor i (since their distance on the metric is 0). However, if doctor j will never propose to hospital A, then hospital A does not want to accept any probability mass from doctor i.

Now, suppose the preferences were fair; if hospital A would want to be matched to doctor j, it would equally want to be matched to doctor i. Thus, hospital A would be able to decide whether to accept or reject doctor i's probability mass independently of whether doctor j will propose to it in a later round or not. Thus, we must allow unfair hospital preferences to achieve the example above, under a proto-metric.

5.3 Stability for General Metrics

Intuitively, a probabilistic allocation is stable if no hospital can offer to some doctors to be matched to it in a way that will improve both the hospital's allocation and those of the doctors. If the hospitals' preferences are not fair, running the Gale-Shapley algorithm fails to output a fair solution for obvious reasons (see Section 5.2). However, even if the hospitals' preferences are fair, finding a non-trivial fair allocation can be quite challenging. This happens because even a mild probabilistic preference for one doctor over another can lead to a situation where a hospital prefers to *always* be matched to one of the doctors and not the other.

Consider the example of two doctors i and j at distance $1/3$, and two hospitals A and B. Suppose both doctors prefer hospital A over hospital B, and hospital A's preferences are

$$\begin{cases} i \succ_A j, & w.p.\ 2/3 \\ j \succ_A i, & w.p.\ 1/3 \end{cases}.$$

Hospital A's preferences are fair. However, in any PIIF allocation, hospital A prefers to be always matched to doctor i, over its outcome in the allocation. This is because if hospital A is always matched to doctor i, it is matched to its first preference with probability $2/3$. However, in any PIIF allocation, hospital A is matched to doctor j with probability at least $1/3$, which implies that it is matched to its first preference with probability no more than $\frac{2}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{1}{3} = \frac{5}{9} < \frac{2}{3}$.

In the above example, A and i were a blocking pair, but the issue was that this alternative is *unfair* (to j). Motivated by this example, and to avoid trivial incompatibilities between fairness and stability, we would like to force the hospitals to make only *fair offers*. However, formalizing such a stability definition that, on the one hand, is strong enough to have a

meaningful guarantee for the hospitals, and on the other hand, is compatible with our fairness requirement (or at least is not trivially incompatible), presents several subtleties. We elaborate on this in the full version of the paper. Instead, we present a minimal (weak) stability definition for general metrics that we use in our negative results (using a weak definition makes the negative results stronger).

5.3.1 Weak Stability

The weak stability definition is guided by simple scenarios. Suppose that the hospitals are A, B and C and that the doctors are i_1, i_2 and j, where doctors i_1 and i_2 are similar (at distance 0) and doctor j is far from them (at distance 1). Consider the following two cases:

- Hospital A is the most preferred hospital by all the doctors, and hospital A prefers doctors i_1 and i_2 over doctor j. It is natural to require that hospital A should be matched to the uniform distribution over i_1 and i_2. Intuitively, this is the "best" allocation for hospital A and doctors i_1 and i_2, subject to fairness. In this case, we maintained fairness by saying that if hospital A is not matched to i_1 and i_2 with probability 1, it is allowed to prefer the alternative allocation of being matched to the uniform distribution over i_1 and i_2 since it is IF.
- Hospital A is the most preferred hospital by doctors i_1 and j, but not by doctor i_2, and hospital A still prefers doctors i_1 and i_2 over doctor j. Suppose that doctor i_2 prefers hospital B over hospital A, that doctor i_2 is matched to hospital B with probability 1, and that the allocation of hospitals A and C has not been determined yet. This time, it is natural to require that hospital A should be matched to doctor i_1 with probability 1. However, if hospital A and doctor i_1 are not matched with probability 1, allowing hospital A to prefer this alternative allocation implies that we allow hospital A to prefer an alternative allocation that does not satisfy IF. We choose to allow this since doctor i_2 prefers its own allocation over being matched to hospital A.

The following definition captures the above intuitions:

▶ **Definition 5** (Weak stability (informal)). *An allocation is (strongly) unstable if there exists a hospital h and an alternative allocation ν over the doctors, such that: (1) The hospital h prefers the alternative allocation ν over its own allocation. (2) Every doctor in the support of the alternative allocation ν prefers hospital h over the hospitals in its support. (3) For every doctor that is not in the support of the alternative allocation ν, either (i) the doctor is at distance 1 from any doctor in the support of the alternative allocation ν or (ii) the doctor prefers every hospital in its own support over the hospital h. (4) For every two doctors in the support of the alternative allocation ν, the allocation ν satisfies IF.*

If there is no such hospital and alternative allocation, then we say that the allocation is weakly stable.

See Definition 17 for formal definition of weak stability.

We note that Definition 3, which is relevant in the proto-metric setting, is a stronger stability. In particular, it implies Definition 5 (when the metric is a proto-metric). In the full proofs in the full version of the paper, we show the barriers for unfair preferences for Definition 5, instead of Definition 3.

5.4 Barriers for General Metrics

In the setting of a general metric and IF hospital preferences, we show that we can make doctors i and j far enough that even under the IF requirement, hospital A's preferences would actually be as described above, i.e., $j \succ_A k \succ_A i \succ_A l$. However, we can make doctors

i and j close enough that an allocation where hospital A is always matched to doctor j, and never to doctor i, would be considered unfair. Then we can arrange the proposals as we did in the case of unfair preferences such that any decision that hospital A makes in the first round can lead to an unfair allocation. See the full version of the paper for details.

6 Open Questions

The new frontier of fairness in two-sided markets raises many fundamental questions for further study. In this work, we present algorithms for finding fair and stable allocations under some restrictions. We show that generalizing this result presents several difficulties. A natural question for further work is either extending the negative results beyond the class of local-proposing algorithms, or finding an algorithm for a more general setting, such as general metrics or unfair preferences.

A possible direction for generalizing the results for general metrics is to have a stronger requirement over the hospital preferences. In the negative results, we use IF preferences for a general metric, and show that effectively they behave similarly to unfair preferences. This indicates that, in the case of a general metric, individual fairness might not be a strong enough fairness requirement over the hospital preferences. We elaborate on this in the full version of the paper.

We present two algorithms in the full version of the paper, which has almost the same guarantees concerning fairness and stability. It is known that different variants of the Gale-Shapley algorithms have different guarantees for optimality and incentive compatibility. In the full version of the paper, we show that doctor-proposing variant) is not optimal for the doctors. For the other variant, we leave this as an open question. We also leave the question of incentive compatibility for future work.

References

1 Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

2 Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

3 Miranda Bogen. All the ways hiring algorithms can introduce bias. *Harvard Business Review*, 6:2019, 2019.

4 Anna Bogomolnaia and Hervé Moulin. A new solution to the random assignment problem. *Journal of Economic theory*, 100(2):295–328, 2001.

5 Slava Bronfman, Avinatan Hassidim, Arnon Afek, Assaf Romm, Rony Shreberk, Ayal Hassidim, and Anda Massler. Assigning israeli medical graduates to internships. *Israel journal of health policy research*, 4(1):1–7, 2015.

6 Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

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

8 Cynthia Dwork and Christina Ilvento. Fairness under composition. *arXiv preprint arXiv:1806.06122*, 2018.

9 Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167*, 2020.

10 Duncan K Foley. *Resource allocation and the public sector*. PhD thesis, Yale University, 1967.

11 Rupert Freeman, Evi Micha, and Nisarg Shah. Two-sided matching meets fair division. *arXiv preprint arXiv:2107.07404*, 2021.

12 David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

13 Ioannis Giannakopoulos, Panagiotis Karras, Dimitrios Tsoumakos, Katerina Doka, and Nectarios Koziris. An equitable solution to the stable marriage problem. In *2015 ieee 27th international conference on tools with artificial intelligence (ictai)*, pages 989–996. IEEE, 2015.

14 Dan Gusfield and Robert W Irving. *The stable marriage problem: structure and algorithms*. MIT press, 1989.

15 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

16 Robert W Irving. Stable marriage and indifference. *Discrete Applied Mathematics*, 48(3):261–272, 1994.

17 Michael P Kim, Aleksandra Korolova, Guy N Rothblum, and Gal Yona. Preference-informed fairness. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

18 S Kiselgof. Matchings with interval order preferences: efficiency vs strategy-proofness. *Procedia Computer Science*, 31:807–813, 2014.

19 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

20 Donald Ervin Knuth. *Stable marriage and its relation to other combinatorial problems: An introduction to the mathematical analysis of algorithms*, volume 10. American Mathematical Soc., 1997.

21 Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

22 Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

23 Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

24 Alvin E Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of political Economy*, 92(6):991–1016, 1984.

25 Alvin E Roth. On the allocation of residents to rural hospitals: a general property of two-sided matching markets. *Econometrica: Journal of the Econometric Society*, pages 425–427, 1986.

26 Alvin E Roth and Marilda Sotomayor. Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541, 1992.

27 Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3082–3092, 2019.

28 Edward G Thurber. Concerning the maximum number of stable matchings in the stable marriage problem. *Discrete Mathematics*, 248(1-3):195–219, 2002.

29 Ariana Tobin. Hud sues facebook over housing discrimination and says the company's algorithms have made the problem worse. *ProPublica (March 28, 2019). Available at https://www.propublica. org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms (last accessed April 29, 2019)*, 2019.

30 Hal Varian. Efficiency, equity and envy. *Journal of Economic Theory*, 9:63–91, 1974.

A Formal Definitions

Let $\mathcal{D}$ be a collection of n doctors and $\mathcal{H}$ be a collection of n hospitals. A *matching* m is a mapping $\mathcal{D}$ to $\mathcal{H}$, i.e., each doctor is mapped to a single hospital. A *probabilistic allocation* π is a mapping from $\mathcal{D}$ to $\Delta(\mathcal{H})$, where for $i \in \mathcal{D}$, $\pi(i)$ represents the *prospect* of doctor i (the probability distribution over hospitals that they receive). Similarly, for a hospital $h \in \mathcal{H}$, $\pi(h)$ represents the *prospect* of doctor h. We focus on the setting where we assign each doctor to a single hospital, so the probabilistic allocation is a distribution over matchings. We assume we are given a *similarity metric* for the doctors $d : \mathcal{D} \times \mathcal{D} \to [0, 1]$.

▶ **Definition 6** (Stochastic Domination). *Let r be a deterministic or probabilistic preference function over a set of individuals (doctors or hospitals) and let p and q be prospects over the same set of individuals. We say that p stochastically dominates q, $p \succeq_r q$, if the following holds:*

$$\forall k \in [n] : \quad \mathop{\mathbf{Pr}}_{o\sim p,r}\left[r^{-1}(o) \leq k\right] \geq \mathop{\mathbf{Pr}}_{o\sim q,r}\left[r^{-1}(o) \leq k\right] \tag{1}$$

A.1 Fairness Definitions

▶ **Definition 7** (Individual Fairness [7]). *An allocation π is individually-fair (IF) with respect to a divergence D, and a similarity metric d, if for all pairs of individuals $i, j \in \mathcal{D}$, the Lipschitz condition $D(\pi(i), \pi(j)) \leq d(i, j)$ is satisfied.*

▶ **Definition 8** (Envy Freeness [10, 30]). *An allocation π is envy-free (EF) with respect to individual preferences $\{\succeq_i\}$ if for all individuals i, for all other individuals j, $\pi(i) \succeq_i \pi(j)$.*

▶ **Definition 9** (Preference-Informed Individual Fairness [17]). *Fix doctors with preferences $r_1, \ldots, r_n$. An allocation π is preference-informed individually fair with respect to a divergence D and a similarity metric d, if and only if for every two doctors i, j, there exists an alternative allocation $p^{i;j}$ such that*

$$D(p^{i;j}, \pi(j)) \leq d(i, j)$$
$$\pi(i) \succeq_i p^{i;j}.$$

▶ **Corollary 10** (PIIF under proto-metrics). *Under proto-metrics we get the following definition: Fix doctors with preferences $r_1, \ldots, r_n$. An allocation π is preference-informed individually fair if for every two doctors i, j, if $d(i, j) = 0$, then either: (i) $\pi(i) = \pi(j)$ or (ii) $\pi(i) \succ_i \pi(j)$.*

▶ **Definition 11** (τ-Preference-Informed Individual Fairness). *An allocation π is τ-PIIF with respect to a similarity metric d, if it is PIIF with respect to the similarity metric d^τ, where d^τ is defined as follows*

$$\forall i, j, \in \mathcal{D} : d^\tau(i, j) = \min\{d(i, j) + \tau, 1\}.$$

▶ **Definition 12** (Strict Individually Fair Preferences). *A set of preferences is strictly individually fair with respect to a proto-metric $d : \mathcal{D} \times \mathcal{D} \to \{0, 1\}$ if for every hospital $h \in \mathcal{H}$:*

- *For every cluster $C \subseteq \mathcal{D}$ and two doctors $i_1, i_2 \in C$, i.e., such that $d(i_1, i_2) = 0$:*

$$\forall r \in [n] : \mathbf{Pr}[r_h(r) = i_1] = \mathbf{Pr}[r_h(r) = i_2].$$

- *For every two clusters $C_1, C_2 \subseteq \mathcal{D}$, either $C_1 \succ_h C_2$ or $C_2 \succ_h C_1$. Where $C_1 \succ_h C_2$ if:*

$$\forall i \in C_1, j \in C_2 : \mathbf{Pr}[r_h^{-1}(i) < r_h^{-1}(j)] = 1.$$

A.2 Stability Definitions

▶ **Definition 13** (Contract). *Given deterministic doctor preferences $\mathcal{P}_D = \{r_i\}_{i\in\mathcal{D}}$, probabilistic hospitals preferences $\mathcal{P}_H = \{r_h\}_{h\in\mathcal{H}}$ and a probabilistic allocation π, a tuple $\mu = (h, i; h', i') \in \mathcal{H}\times\mathcal{D}\times\mathcal{H}\times\mathcal{D}$ is a contract if $d(i, i') = 1$, $\pi^\mu(h) \succ_h \pi(h)$ and $r_i(h) < r_i(h')$, where:*

$$\forall j \in \mathcal{D}\backslash\{i, i'\} : \pi^\mu(j) = \pi(j) \tag{2}$$

$$\pi^\mu(i) = \begin{cases} h, & \pi(i) = h' \wedge \pi(i') = h \\ \pi(i), & \text{otherwise} \end{cases} \tag{3}$$

$$\pi^\mu(i') = \begin{cases} h', & \pi(i) = h' \wedge \pi(i') = h \\ \pi(i'), & \text{otherwise} \end{cases} \tag{4}$$

▶ **Definition 14** (Active Contract). *Given deterministic doctor preferences $\mathcal{P}_D = \{r_i\}_{i\in\mathcal{D}}$, probabilistic hospitals preferences $\mathcal{P}_H = \{r_h\}_{h\in\mathcal{H}}$ and a probabilistic allocation π, a contract $\mu = (h, i; h', i') \in \mathcal{H}\times\mathcal{D}\times\mathcal{H}\times\mathcal{D}$ is an active contract if*

$$\mathbf{Pr}[\pi(h) = i' \wedge \pi(i) = h'] > 0.$$

▶ **Definition 15** (Contract Stability). *The allocation π is contract stable if there are no active contracts.*

▶ **Definition 16** (τ-Contract Stability). *Denote by $S_\pi \subseteq \mathcal{D}\times\mathcal{H}\times\mathcal{D}\times\mathcal{H}$ the set of all active contracts in an allocation π. An allocation π is τ-singleton contract stable if*

$$\mathbf{Pr}[\bigvee_{(h,i;h',i')\in S_\pi} (\pi(h) = i' \wedge \pi(h') = i)] \leq \tau.$$

▶ **Definition 17** (Weak Ex-Ante Stability). *Let π be a probabilistic allocation, $h \in \mathcal{H}$ be a hospital, $\mathcal{D}^* \subseteq \mathcal{D}$ be a set of doctors and $\sigma \in \Delta(\mathcal{D}^*)$ be a distribution. We say that $\nu = (h, \mathcal{D}^*, \sigma)$ is a selectively fair alternative allocation if σ satisfies:*

- *For every two doctors $i, j \in \mathcal{D}^*$, the distribution σ is individually fair with respect to i and j.*
- *For every doctor $i \in \mathcal{D}\backslash\mathcal{D}^*$, if there exists a doctor $j \in \mathcal{D}^*$ such that $d(i, j) < 1$, then for every hospital $h' \in supp(\pi(i))$, $h' \succ_i h$.*
- *For every $i \in \mathcal{D}^*$ and $h' \in supp(\pi(i))$, $h \succeq_i h'$.*

We say that ν is active if $\sigma \succ_h \pi(h)$. We say that the allocation π is weakly ex-ante stable if there are no active selectively fair alternative allocations.