Interview Hoarding*

Vikram Manjunath
University of Ottawa
vikramma@gmail.com

Thayer Morrill
North Carolina State University
thayer_morrill@ncsu.edu

May 11, 2021

Abstract

Many centralized matching markets are preceded by interviews between the participants. We study the impact on the final match of an increase in the number of interviews for one side of the market. Our motivation is the match between residents and hospitals where, due to the COVID-19 pandemic, interviews for the 2020-21 season of the National Residency Matching Program were switched to a virtual format. This drastically reduced the cost to applicants of accepting interview invitations. However, the reduction in cost was not symmetric since applicants, not programs, previously bore most of the costs of in-person interviews. We show that, starting from a situation where the final matching is stable, if doctors can accept more interviews, but the hospitals do not increase the number of interviews they offer, then no doctor is better off and many doctors are potentially harmed. This adverse consequence is the result of what we call *interview hoarding.* We prove this

*We thank Anna Sorensen and Alkas Baybas for raising the question that sparked this paper. We also thank Alex Chan, Adrienne Quirouet, Assaf Romm, Al Roth, Erling Skancke, Colin Sullivan, William Thomson, and seminar audiences at North Carolina State University, Stanford, University of Arizona, and University of Lausanne for helpful comments and discussions.

analytically and characterize optimal mitigation strategies for special cases. We use simulations to extend these insights to more general settings.

Keywords: NRMP, Deferred acceptance, Interviews, Hoarding

1 Introduction

Perhaps the most well-known application of matching theory is the entry-level labor market for physicians. In 2021, 37,470 positions were matched through the National Resident Matching Program (NRMP). The matching process consists of two steps. First, each physician interviews with a set of residency programs. Second, programs and physicians submit rank-order lists of those they interview to a centralized clearinghouse. This clearinghouse, run by the NRMP, matches physicians to residency programs using a version of Gale and Shapley's (1962) Deferred Acceptance (DA) Algorithm (Roth and Peranson, 1999).

In practice, both programs and applicants are constrained in the number of interviews they can take part in. Prior to the COVID-19 pandemic, interviews were done in person. These interviews were particularly costly for physicians since they not only had to bear travel expenses but also had to take days off from clinical rotations. The cost to programs was mainly in terms of time. For the 2020-21 matching season, interviews were conducted virtually. While this dramatically decreased the cost of interviews for physicians, it did not change the costs for the programs. We are interested in the implications of this asymmetric change on the eventual match.

We follow the approach of Echenique et al. (2020) and assume complete information about preferences and non-strategic offers and acceptance of interviews. In reality, interviews serve the obvious purpose of preference formation. However, in the context of the NRMP match, they are also important in coordinating mutual rankings across the two sides. Most specialties have hundreds of programs and thousands of candidates, so the rank-order lists submitted to the NRMP are necessarily partial. Moreover, a program and a candidate can only match if they rank

one another. This coordination across the two sides is important independent of preference formation. The complete information set up of Echenique et al. (2020) keeps the analysis focused on bottlenecks that arise in the interview phase.[1] The assumption of non-strategic behavior makes the analysis tractable. These modeling choices allow us to identify a subtle bottleneck caused by interviews that would likely be lost in the analysis of a more complex model.

The focus of our study is the effects of doctors accepting more interview invitations without a corresponding increase to the number of invitations extended by programs. If some doctors accept more interviews, and if the total number of invitations extended does not change, then some of the doctors necessarily receive fewer invitations. It seems natural to intuit that at least doctors with more interviews benefit from the lower costs, even if those with fewer interviews are harmed. However, for any market where the eventual matching is stable—which we would expect at the steady state of a well-functioning market—we show a surprising result: No physician is better off when more interviews can be accepted.

The intuition for this result is as follows. Consider a highly sought after physician: one who is offered interviews at the leading programs and ends up matched with her favorite program. When interviews become cheaper, she will accept more interviews.[2] However, as she would already have matched with her favorite program, the interviews she accepts are from inferior programs. These interviews do not help her: she ultimately matches with the same program as before. The interviews are, in effect, wasted. We refer to this as *interview hoarding*. Interview hoarding has a cascading affect. The physicians who otherwise would have filled these wasted interview slots now interview with programs they consider inferior. These physicians may have more interviews, but they do not have better inter-

[1] We weaken these assumptions, to an extent, in the appendices. Analogs of some of our results hold for a version of our model that includes preference formation. Simulation results with more sophisticated heuristic based interview offers are consistent with our results as well.

[2] While we have assumed that agents have complete information, if there were an arbitrarily small amount of uncertainty about others' preferences, she would accept additional costless interviews.

views in a precise sense: the doctor rates every new interview as worse than the program she matched with before. Physicians are ultimately divided into three categories: physicians who hoard interviews worse than their eventual match; physicians who receive more but worse interviews; and physicians who receive fewer and worse interviews. The first category is indifferent between the new costs and the old. The latter two categories are harmed under the new cost. Thus, when physicians accept more interviews but programs do not react, the ultimate match is Pareto inferior from the physicians' perspective.[3]

Having shown that increases to doctors' abilities to accept interviews has adverse welfare consequences, we turn to mitigation policies. We consider policies that limit the numbers of interviews that programs can offer and candidates can accept. Though there are essentially no such policies that *always* (for every preference profile) yield a stable final matching (Proposition 1), we characterize such policies for "common preferences" (Proposition 2). These are salient preference profiles where every doctor ranks the programs the same way and every program ranks the doctors the same way. The policies we characterize are such that there is a common cap on the number of interviews any program can offer or any candidate can accept. We also show that if the programs' interview capacities are fixed, say at l, then the number of blocking pairs increases and the match rate decreases as the doctors' interview cap moves further away from l in either direction (Proposition 3).

Our analytical results can advise policies for more general settings where preferences are not quite common, but have a common component. We use simulations to show that the lessons from our analytical results hold up under weaker assumptions. Though the optimal cap on doctors' interview capacities depends on the parameters of the model—and in practice would have to be determined empirically—our simulations indicate that it is no higher than the number of interviews that the programs offer.[4]

[3]The example in Section 1.2 demonstrates that there need not exist a Pareto ranking from the programs' perspective.

[4]This is true whether we define the optimality of a policy as maximizing the expected proportion

1.1 Related Literature

While there is a large literature on the post-interview NRMP match,[5] there are relatively few papers that incorporate the pre-match interview process. One of the first to explicitly model interviews in the classic one-to-one matching model is Lee and Schwarz (2017). In this model, before participating in a centralized, two-sided match, firms learn their preferences over workers by engaging in costly interviews. They show that even if firms and workers interview with exactly the same numbers of agents, the extent of unemployment in the final match depends critically on the overlap between the sets of workers that firms interview. Three other recent papers that incorporate pre-match interviews include Kadam (2021), Beyhaghi (2019), and Echenique et al. (2020).

Like our paper, Kadam (2021) considers the implications of loosened interview constraints for doctors. However, the focus is on the strategic allocation of scarce interview slots. For the sake of tractability, the analysis is for a stylized model of large markets. Under the assumption of common preferences over programs, he shows that increasing doctors' capacities may increase total surplus, but not in a Pareto-improving way. Moreover, the match rate decreases. He also highlights that when preferences are not necessarily common, the effect is ambiguous, since increased interview capacities dilute doctors' signaling ability.

Beyhaghi (2019) also performs a strategic analysis of a stylized large market model. However, she considers a slightly different set up with *application* caps for doctors and interview caps for programs. While similar, application caps are not exactly the same as interview caps: they constrain the number of programs a doctor can express interest in at the outset of the interview matching phase, but not the number of interviews she can accept at the end. In her model, inequity in the application caps decreases the expected total surplus. Moreover, when interview capacity is low, low application caps are socially desirable.

In our model, the agents do not choose interviews strategically. Determining

of positions that are filled or minimizing the expected number of blocking pairs.

[5] See the multitude of papers following Roth and Peranson (1999).

the optimal set of interviews is closely related to the portfolio choice problem of Chade and Smith (2006). They solve for the optimal portfolio when an agent chooses a portfolio of costly, stochastic options, but only consumes one of their realizations. In order to apply the optimal solution to the interview scheduling problem, one would have to pin down precisely the probability of any given pair matching. This is what makes strategic analysis of the problem intractable without severe simplifying assumptions (such as those in the papers we have mentioned above).

As in Echenique et al. (2020), which is methodologically closest to ours, we sidestep this issue. They explain a puzzling empirical pattern resulting from the NRMP match: 46.3% of the physicians were matched to their top ranked residency programs and 71.1% where matched to a program they ranked in their top three. These statistics seem to contradict surveys indicating that many doctors have similar preferences over residency programs. They provide an explanation for this phenomenon by pointing out the importance of the interviewing process that precedes the match. Roughly speaking, the pre-match interviewing process restricts the preferences that the physicians actually submit to the NRMP. Therefore, a proper interpretation is not that the physicians matched with their most preferred programs but rather that they matched with their most preferred programs *among those they interviewed with*.

Our work is complementary with these papers in the sense that they highlight the importance of understanding the prematch interviews for properly evaluating the NRMP match itself.

1.2 Motivating Example

We present the intuition behind the welfare loss from increased interview capacity for doctors with a simple example. Consider a market with four doctors $\{d_1, \dots, d_4\}$ and four hospitals $\{h_1, \dots, h_4\}$. The agents' preferences are as follows:

d_1	d_2	d_3	d_4	h_1	h_2	h_3	h_4
h_4	h_2	h_2	h_2	d_4	d_1	d_2	d_4
h_1	h_3	h_1	h_3	d_1	d_2	d_1	d_3
h_2	h_1	h_3	h_4	d_2	d_4	d_3	d_2
h_3	h_4	h_4	h_1	d_3	d_3	d_4	d_1

Suppose that the interview capacities of the doctors and hospitals are:

d_1	d_2	d_3	d_4	h_1	h_2	h_3	h_4
1	2	1	1	1	1	2	1

Interviews are initially offered by hospitals: h_1 invites d_4, h_2 invites d_1, h_3 invites d_1 and d_2, and h_4 invites d_4. As d_1 can accept only one invitation, she turns h_3 down. Since d_4 can only accept one invitation, she declines h_1's invitation. Doctor d_2 holds on to her invitation from h_3. Hospitals h_1 and h_3 then offer interviews to d_1 and d_3, respectively. Doctor d_1 rejects her invitation from h_2. After h_2 invites d_2, the final interviews are:

d_1	d_2	d_3	d_4
h_1	$\{h_2, h_3\}$	h_3	h_4

The final matching is computed by applying the doctor-proposing Deferred Acceptance algorithm to the agent preferences (restricted to agents they interview with). The outcome is therefore:

d_1	d_2	d_3	d_4
h_1	h_2	h_3	h_4

The market functions well in the sense that the final outcome is stable with regard to the actual (as opposed to restricted) preferences.

Now suppose d_1 is able to accept an additional interview (and all other interview capacities remain the same). In this case she does not reject h_2's invitation. The interview schedule is:

d_1	d_2	d_3	d_4
$\{h_1, h_2\}$	h_3	h_3	h_4

This leads to the final matching:

d_1	d_2	d_3	d_4
h_1	h_3		h_4

Doctor d_1 does not benefit from the additional interview. The interview she adds is with a hospital that she finds worse than her original match. However, her acceptance of h_2's invitation comes at the expense of both d_2 and d_3: both now receive worse assignments. In fact, the final matching is no longer stable. That none of the doctors are better off is not unique to this example—we show that this is generally true (Theorem 1). The programs, however, are not unanimously better or worse off: h_3 is better off while h_2 is worse off.

2 The Model

A **market** consists of a triple (D,H,P), where D is a finite set of **doctors**, H is a finite set of **hospitals**, and P is a profile of strict **preferences** for the doctors and hospitals. We assume that there are at least two doctors and two hospitals: $|D| \geq 2$ and $|H| \geq 2$. For each $h \in H$, $\mathcal{P}_h$ is the set of strict preferences over $D \cup \{h\}$, and for each $d \in D$, $\mathcal{P}_d$ is the set of strict preferences over $H \cup \{d\}$. The set of preference profiles is $\mathcal{P} \equiv \times_{i \in H \cup D} \mathcal{P}_i$.

There are two phases to the matching process. The first is a decentralized interview phase and the second is the centralized matching phase. The former involves many-to-many matching while the latter is a standard one-to-one matching problem (Roth and Sotomayor, 1990).

A *many-to-many* matching is a function $\nu : H \cup D \to 2^{H \cup D}$ such that, for each $d \in D$ and $h \in H$, $\nu(d) \subseteq H$, $\nu(h) \subseteq D$, and $h \in \nu(d)$ if and only if $d \in \nu(h)$.

For each $h \in H$, let $\iota_h \in \mathbb{N}$ be **h's interview capacity**. Similarly, for each $d \in D$, let $\kappa_d \in \mathbb{N}$ be **d's interview capacity**. We call the profile $(\iota,\kappa) = ((\iota_h)_{h \in H}(\kappa_d)_{d \in D})$ the **interview arrangement**. An **interview matching** is a many-to-many matching ν such that for every doctor d, $|\nu(d)| \leq \kappa_d$ and for every hospital h, $|\nu(h)| \leq \iota_h$.

An interview matching ν is **pairwise stable** if there is no doctor-hospital pair (d,h) such that $h \notin \nu(d)$ but:

- either $|\nu(h)| < \iota_h$ and $d \, P_h \, h$ or there exists a $d' \in \nu(h)$ such that $d \, P_h \, d'$, and
- either $|\nu(d)| < \kappa_d$ and $h \, P_d \, d$ or there exists a $h' \in \nu(d)$ such that $h \, P_d \, h'$.

A matching is a function $\mu : H \cup D \to H \cup D$ such that $\mu(h) \in D \cup \{h\}$, $\mu(d) \in H \cup \{d\}$, and $\mu(d) = h$ if and only if $\mu(h) = d$. We say that (d,h) is a **blocking pair** of matching μ if $h \, P_d \, \mu(d)$ and $d \, P_h \, \mu(h)$. A matching is **stable** if it does not have a blocking pair.

To describe how the market works, we follow the approach of Echenique et al. (2020) by assuming complete information and non-strategic behavior.[6] This means that hospitals naïvely make offers to their most preferred doctors and these offers, if rejected, trickle down to less preferred doctors. Thus, given (ι,κ) and $P \in \mathcal{P}$, the final matching, which we call the **(ι,κ)-matching**, is the outcome of the following two phase process:[7]

Phase 1: The interview matching ν is the hospital-optimal pairwise stable many-to-many matching where the capacities of the hospitals and doctors are given by ι and κ, respectively. This can be computed by applying the hospital-proposing deferred acceptance (DA) algorithm: each $h \in H$ is matched with up to ι_h doctors and each $d \in D$ is matched with up to κ_d hospitals. Since we ignore the informational aspect of the problem, the input to DA is a choice function for each agent that is responsive to her preference relation and constrained by her interview capacity.[8] The hospital-proposing DA algorithm is

[6] In Appendix A, we discuss a version of our model and some of our results when preferences are formed during the interviews.

[7] We only differ from Echenique et al. (2020) in that we set the interview matching to be the *hospital-optimal* many-to-many stable matching, while they set it to be the *doctor-optimal* one. Their choice is appropriate for the question they ask. However, we have chosen to approximate the interview phase through the hospital-proposing DA. This difference does not drive our results, as we explain in Appendix B.

[8] For the sake of completeness, we define in Appendix C the acceptant and responsive choice

an approximation of the decentralized process by which hospitals invite doctors, extending invitations to further doctors when invitations are declined.

Phase 2: The (ι, κ)-matching is chosen by doctor-proposing DA. The input to DA is the true preference profile restricted to the interview match, $(P_i|_{\nu(i)})_{i \in D \cup H}$.

Given $P \in \mathcal{P}$, we say that **(ι, κ) is adequate** if the (ι, κ)-matching at P is stable. We interpret (ι, κ) being adequate at a profile P as a sign that the market is functioning well. Otherwise, a blocking pair could alter their behavior to improve their lot. In other words, using stability as our notion of equilibrium, (ι, κ) being adequate is equivalent to the market being in equilibrium. If (ι, κ) is adequate at each $P \in \mathcal{P}$, then we say that (ι, κ) is **globally adequate**.

Finally, we define a welfare comparison between matchings. Given a pair of matchings μ and μ', we say that **no doctor prefers μ' to μ** if, for each $d \in D$, $\mu(d) \, R_d \, \mu'(d)$.

3 Welfare Impact of Increased Interviews

Our aim is to study how a change in the interview costs impacts a market. We consider a market that is initially at equilibrium. Starting from such an equilibrium, the goal is to understand the welfare consequences of a shock that permits doctors to accept more interviews. That is, starting with $P \in \mathcal{P}$ and (ι, κ) that is adequate at P, we consider an increase in the doctors' interview capacities to κ' and compare the (ι, κ)-matching to the (ι, κ')-matching.

The doctor who accepted more interviews in our example in Section 1.2 did not benefit from it. Our main result shows that this is true in general.

functions that we appeal to while running DA to compute the interview matching. According to these choice functions, hospitals naïvely select their most preferred doctors from each set. In Appendix D, we consider alternative choice functions that reflect common heuristics. Simulations involving such a heuristic based choice are closer to the results of the 2021 match, but the qualitative effects are similar. The simpler assumption of naïve behavior renders the results more transparent.

Theorem 1. *Starting at an adequate arrangement, doctors do not benefit from increases to their interview capacities. That is, if* (ι, κ) *is adequate at* P *and* κ' *is such that, for each* $d \in D$, $\kappa'_d \geq \kappa_d$, *then the no doctor prefers the* (ι, κ')*-matching to the* (ι, κ)*-matching.*

Proof. Let ν and μ be the interview and final matchings, respectively, under (ι, κ). Similarly, let ν' and μ' be the interview and final matchings under (ι, κ'). We frame the temporal language below in reference to a hypothetical change in doctors' interview capacities from κ ("before") to κ' ("after").

We first establish a number of properties of the interview matchings. The intuition for these results comes from one of the classical results in two-sided matching theory: When the set of men increases, no man benefits from this increased competition while no woman is harmed.[9] In our setting, an increase in the number of interviews a doctor can participate in plays the role of additional men participating in the market. This means that the hospitals are able to interview better doctors. However, there is a tension between interviewing better doctors and interviewing the "right" doctors. Thus improving the set of candidates a hospital interviews does not necessarily translate to an improvement in its eventual match. Those doctors whose capacities do not change do not see an improvement in the hospitals that interview them. For those whose capacities do increase, the stability of the original matching means that none of the new prospects come to fruition.

Lemma 1. *No doctor rejects a hospital she previously interviewed with.*

Proof. Suppose not. In the interview matching phase (under capacities κ'), let d be the first doctor to reject a hospital h that she interviewed with under capacities κ. As d has at least as much interview capacity, she must have received a new proposal from some hospital h'. As h' did not propose to d before, it must have been rejected by some doctor $d' \in \nu(h)$, a doctor it previously interviewed. But this contradicts d being the first doctor to reject a hospital she previously interviewed with. □

[9] See Theorem 2.25 of Roth and Sotomayor (1990).

We cannot say whether a doctor prefers her interviews under κ versus κ' as we only have a doctor's preferences over individual hospitals and not sets of hospitals. However, we show—in a specific sense—that while a doctor may get new interviews, she does not get better interviews.

Lemma 2. *No doctor has a new interview better than her previous matching: if $h \in \nu'(d) \setminus \nu(d)$, then $\mu(d) \, P_d \, h$.*

Proof. Suppose not. Let d be the first doctor when DA is run during the interview phase under capacities κ' to receive a proposal from a hospital $h \notin \nu(d)$ such that $h \, P_d \, \mu(d)$. As h did not previously propose to d, h must have been rejected by a doctor that it previously interviewed. This contradicts Lemma 1. □

In the classical result, no man benefits from the increased competition due to additional men and also no woman is harmed. An analogous result holds in our framework. A hospital either has the same set of interviews, additional interviews, or it interviews new doctors it prefers to her previous interviews. In any scenario, the hospital's set of interviews (weakly) improves.

Lemma 3. *Suppose a hospital h interviews a doctor d under κ'. If h previously interviewed d' and prefers d' to d, then h continues to interview d': if $d \in \nu'(h)$, $d' \in \nu(h)$, and $d' \, P_h \, d$, then $d' \in \nu'(h)$.*

Proof. As $d' \, P_h \, d$, h proposes to d' before it proposes to d when DA is run in the interview phase under κ'. By Lemma 1, h is not rejected by any doctor it previously interviewed. As h proposes to d under κ', it must have already proposed to but not have been rejected by d'. Therefore, h continues to interview d'. □

To complete the proof of Theorem 1, we show that if a doctor is rejected by a hospital during the matching phase under κ, then she is not matched to that hospital under κ'. We proceed by induction on the round (of DA in the matching phase under κ) in which the doctor was rejected, and our inductive hypothesis is that if doctor d was rejected by hospital h in round k under κ, then under κ', either she no longer interviews with h or she is rejected in round k or earlier.

For the base step, consider a doctor d who was rejected by hospital h in the first round under κ, and let d' be the doctor that h tentatively accepts. If d does not interview with h under κ' ($d \notin \nu'(h)$), then we are done. Therefore, suppose $d \in \nu'(h)$. By Lemma 3, since h prefers d' to d and it interviews d, it also interviews d' ($d' \in \nu'(h)$). Doctor d' does not have any new interviews with a hospital she prefers to h since $h\, R_{d'}\, \mu(d')$ and by Lemma 2 she does not get a new interview with a hospital she prefers to $\mu(d')$. Therefore, d' continues to propose to h in the first round even under the new capacities and d continues to be rejected by h in favor of d' or possibly a doctor h prefers even more.

To complete the inductive argument, suppose that doctor d was rejected by hospital h in favor of doctor d' in round k under κ. If $h \notin \nu'(d)$, then we are done. Otherwise, again by Lemma 3, $d' \in \nu'(h)$. Under κ, d' proposes to h in round k or earlier. Therefore, d' was rejected by all hospitals she interviewed with and prefers to h in an earlier round. By the inductive hypothesis, for any hospital h' that rejected d' under κ, either d' no longer interviews with h' or h' has already rejected d' by round k under κ'. Therefore, under κ', d' proposes to h in either round k or a previous round. In either case, by round k, under κ', h has already received a proposal it prefers to d. Therefore, doctor d is rejected by hospital h under κ' in round k or earlier.

This shows that if d was rejected by hospital h under the old capacities, then d is not matched to h under the new capacities. Note that d has no new interviews with a hospital she prefers to $\mu(d)$. Therefore, if $h\, P_d\, \mu(d)$ and $h \in \nu'(d)$, then $h \in \nu(d)$ and h rejected d in some round under the old capacities. Therefore, h also rejects d under the new capacities. In particular, under the new capacities, d is not matched to a hospital she prefers to $\mu(d)$. □

Theorem 1 tells us that doctors increasing the number of interviews they accept will either have no impact on the resulting matching or will make the new matching Pareto worse from the doctors' perspective. The example in Section 1.2 illustrates that there are instances where increasing the interview capacity does result in a Pareto inferior outcome. This example is not pathological. Lemmas 1 and 2

highlight the root cause of the inferior match, which is interview hoarding.

Under our simplifying assumptions that agents are non-strategic and have complete information, Theorem 1 implies that the shift to virtual interviews for the 2020-21 season of the NRMP ought to have led to an inferior matching. In Appendix D, we contrast simulation results for this naïve behavior with the common heuristic of including a "safety" candidate when choosing a set. While the NRMP has touted the high match rate for 2021, this may be driven by hospitals being matched to safety candidates (under such heuristic behavior) rather than being unmatched (under naïve behavior). Other than the match rate, heuristic choice does not qualitatively affect the results.

4 Adequate Arrangements

Theorem 1 assumes that the initial profile of interviews was adequate in the sense that the outcome of the two-phase process is a stable matching. We interpret this assumption as a characteristic of a well-functioning market in steady state equilibrium. A natural question is how many interviews need to take place and what does the distribution of interviews need to be in order for an interview profile to be adequate. Of course, in general, the answer will depend on specifics of the market, such as the ratio of doctors to hospitals and how correlated or aligned preferences are. However, we are able to provide tight characterizations for certain "end-point" cases that provide intuition for more general markets.

4.1 Globally Adequate Arrangements

In studying adequate arrangements, we first discuss worst case performance: what arrangements are adequate *for every* preference profile? It turns out that only very extreme arrangements satisfy this property. We characterize these arrangements in our next result.

Proposition 1. *Arrangement* (ι, κ) *is globally adequate, if and only if either*

1. *every doctor and every hospital has only unit interview capacity—that is, for each* $d \in D, \kappa_d = 1$ *and for each* $h \in H, \iota_h = 1$*—or*

2. *every doctor and every hospital has high interview capacity—that is, for each* $d \in D, \kappa_d \geq \min\{|D|, |H|\}$ *and for each* $h \in H, \iota_h \geq \min\{|D|, |H|\}$.

Proof. We first prove necessity. Suppose that (ι, κ) is globally adequate.

We start by establishing that if one doctor or hospital has greater than unit interview capacity, then every doctor and hospital has interview capacity of at least two. Stated differently, if any doctor or hospital has unit capacity, then all doctors and hospitals have unit capacity. We denote by ν the interview matching and by μ the (ι, κ)-matching.

Claim 1. *1. If there is* $d \in D$ *such that* $\kappa_d > 1$*, then for each* $d' \in D, \kappa_d \geq 2$ *and for each* $h \in H, \iota_h \geq 2$*, and*

2. If there is $h \in H$ *such that* $\iota_h > 1$*, then for each* $h' \in H, \iota_h \geq 2$ *and for each* $d \in D, \kappa_d \geq 2$.

Proof. We prove only the first statement as the proof of the second statement is analogous—it requires only a reversal of the roles of doctors and hospitals.

Suppose, for the sake of contradiction, there is a globally adequate (ι, κ) where there exists a $d_1 \in D$ such that $\kappa_{d_1} > 1$ and a $h_2 \in H$ such that $\iota_{h_2} = 1$. Let $h_1 \in H \setminus \{h_2\}$ and $d_2 \in D \setminus \{d_1\}$. Consider $P \in \mathcal{P}$ where each doctor ranks h_1 first and h_2 second, and each hospital ranks d_1 first and d_2 second. All hospitals offer an interview to d_1 and as $\kappa_{d_1} > 1$, d_1 accepts interviews from at least h_1 and h_2. Since $\iota_{h_2} = 1$, h_2 only interviews d_1. Let μ be the (ι, κ)-matching. Since (ι, κ) is adequate, μ is stable, so $\mu(d_1) = h_1$, as h_1 and d_1 are mutual favorites. Therefore, $\mu(h_2) = h_2$ as h_2 only interviews d_1. Note that (d_2, h_2) forms a blocking pair of μ as $h_2 \, P_{d_2} \, \mu(d_2)$, since $\mu(d_2) \notin \{h_1, h_2\}$, and $d_2 \, P_{h_2} \, h_2$. This contradicts the stability of μ and thus the assumption that (ι, κ) is globally adequate. We have therefore established that if there is $d \in D$ such that $\kappa_d > 1$, then for each $h \in H$, $\iota_h \geq 2$.

We now prove that if there is a $d_1 \in D$ such that $\kappa_{d_1} > 1$, then for each $d \in D$, $\kappa_d \geq 2$. Suppose for the sake of contradiction that there is $d_2 \in D$ such that

$\kappa_{d_2} = 1$. Let $h_1, h_2 \in H$. Consider $P \in \mathcal{P}$ such that each doctor ranks h_1 first and h_2 second, and each hospital ranks d_1 first and d_2 second. As we have shown above, $\iota_{h_1}, \iota_{h_2} \geq 2$, so both h_1 and h_2 offer interviews to both d_1 and d_2. Since h_1 is her favorite hospital, d_2 accepts its offer. Thus, $\nu(d_2) = \{h_1\}$. However, $\mu(d_1) = h_1$ since d_1 and h_1 are mutual favorites, so $\mu(d_2) = d_2$. This means that (d_2, h_2) form a blocking pair of μ as the only hospital d_2 prefers to h_2 is h_1. This contradicts the stability of μ and thus the assumption that (ι, κ) is globally adequate. □

We complete the proof of necessity by showing that neither a doctor nor a hospital can have an intermediate capacity.

Claim 2. *There is no $d \in D$ such that $1 < \kappa_d < \min\{|D|, |H|\}$, and there is no hospital h such that $1 < \iota_h < \min\{|D|, |H|\}$.*

Proof. We prove this statement for the case where $|D| \leq |H|$. The proof when $|H| < |D|$ is symmetric.

Suppose for the sake of contradiction that $d_1 \in D$ is such that $\kappa_{d_1} = k$ where $1 < k < |D|$. Let $P \in \mathcal{P}$ be such that for i from 1 through $k+1$:

$$P_{d_1} : h_2, h_3, \ldots, h_{k+1}, h_1, \ldots$$
$$P_{h_i} : h_i, h_1, \ldots, h_{i-1}, h_{i+1}, \ldots$$
$$P_{h_1} : d_1, d_2, \ldots$$
$$P_{h_i} : d_i, d_1, \ldots, d_{i-1}, d_{i+1}, \ldots$$

We have constructed the preference profile P such that:

- For each i from 1 through $k+1$, d_i and h_i are matched in every stable matching.
- Each of the $k+1$ hospitals $h_1, \ldots, h_{k+1}$ offers d_1 an interview.
- Doctor d_1 accepts interview offers from hospitals $h_2, \ldots, h_{k+1}$, but not from h_1.

The first and third points are immediate consequences of the preferences. The second is a consequence of the first part of Claim 1: since $\kappa_{d_1} > 1$, every hospital

has an interview capacity of at least two and ranks d_1 in its top two. However, this contradicts the definition of μ as the (ι,κ)-matching, since $h_1 \notin \nu(d_1)$ yet by stability, $h_1 = \mu(d_1)$.

A similar construction shows that there is no $h \in H$ such that $1 < \iota_h < |D|$. Suppose for the sake of contradiction that $h_1 \in H$ is such that $\iota_{h_1} = l$ where $1 < l < |D|$. Let $P \in \mathcal{P}$ be such that for i from 1 through $l+1$:

$$
\begin{aligned}
P_{d_1} &: h_1, h_2, \ldots \\
P_{d_i} &: h_i, h_1, \ldots, h_{i-1}, h_{i+1}, \ldots \\
P_{h_1} &: d_2, d_3, \ldots, d_{l+1}, d_1 \\
P_{h_i} &: d_i, d_1, \ldots, d_{i-1}, d_{i+1}, \ldots
\end{aligned}
$$

By the second part of Claim 1, since $\iota_{h_1} > 1$, every doctor has a capacity of at least two. Therefore:

- For each i from 1 through $l+1$, d_i and h_i are matched in every stable matching.
- Each of the l doctors $d_2, \ldots, d_{k+1}$ accepts an interview from h_1.
- Hospital h_1 does not offer d_1 an interview.

Thus, $h_1 \notin \nu(d_1)$, so $h_1 \neq \mu(d_1)$. This contradicts the stability of μ, the (ι,κ)-matching, and in turn the assumption that (ι,κ) is globally adequate. □

We now turn to sufficiency. If every agent has an interview capacity of one, then the interview matching is actually a matching. Moreover, it is a stable matching. So, suppose that each agent has an interview capacity of at least $\min\{|D|,|H|\}$. If $|D| = |H|$, then the interview matching involves an interview between every mutually acceptable doctor-hospital pair. This means that the (ι,κ)-matching is the doctor optimal stable matching under unrestricted preferences, which is stable. We now show, that even if $|D| < |H|$ or $|D| > |H|$, the (ι,κ)-matching, μ, is stable. Suppose the doctor-hospital pair (d,h) blocks μ. By defintion of μ as the (ι,κ)-matching, if $h\, P_d\, \mu(d)$ and $d\, P_h\, \mu(h)$, then $h \notin \nu(d)$.

Suppose $|D| < |H|$. Since $\iota_h \geq |D|$, h would have offered an interview to d and would have been rejected during the interview matching phase, so $\nu(d)$ contains κ_d hospitals that d prefers to h. Since $h \, P_d \, \mu(d)$, and $\mu(d) \in \nu(d) \cup \{d\}$, this means $\mu(d) = d$. Then, d is rejected by every hospital in $\nu(d)$ during the application of DA in the matching phase. However, $|\nu(d)| = \kappa_d \geq |D|$ and since d is acceptable to every hospital in $\nu(d)$, she is only rejected when another doctor applies. However, this implies that when DA terminates in the matching phase, every hospital in $\nu(d)$ has tentatively accepted some doctor other than d, which is a contradiction—there are not enough such doctors.

Suppose $|H| < |D|$. Since $\kappa_d \geq |H|$, d does not reject any interviews she is offered. Since $h \notin \nu(d)$, h offers interviews to and has them accepted by $\iota_h \geq |H|$ doctors whom it prefers to d. Since $d \, P_h \, \mu(h)$, h does not receive a proposal from any $d' \in \nu(h)$ during the application of DA in the matching phase since it finds all such d' better than d. This implies that each $d' \in \nu(h)$ is tentatively accepted by some hospital other than h when DA terminates, which is a contradiction—there are not enough such hospitals. □

Proposition 1 highlights a previously overlooked role that the interview phase plays in determining whether or not the ultimate NRMP match is stable. While interviews are necessary for agents to gain information, we learn from Proposition 1 that interviews can also act as a bottleneck. Even with complete information, once any agent is capable of participating in more than one interview, all agents must interview with essentially the entire market to be certain that the ultimate match is stable.

4.2 Homogeneous Arrangements

The distribution of interviews is an essential factor in the stability of the NRMP match. So, it is natural to consider market interventions when the interview phase is out of balance. Our motivating question concerns what happens when there is an increase to the number of interviews doctors can accept. The most straight-

forward intervention is to cap the number of interviews an agent can participate in. Here we consider a **homogeneous arrangement**: all doctors face the same cap and all hospitals face the same cap, but we allow the doctor and hospital caps to potentially differ. In other words, the arrangement would be described by two numbers: an interview capacity $l \in \mathbb{N}$ for hospitals and an interview capacity $k \in \mathbb{N}$ for doctors. The pair (l,k) corresponds to the arrangement (ι,κ) where for each $h \in H, \iota_h = l$ and for each $d \in D, \kappa_d = k$.

By Proposition 1 a homogenous arrangement (l,k) can only be globally adequate if $l = k = 1$ or $l,k \geq \min\{|D|,|H|\}$. Nonetheless, (l,k) *may* be adequate for a specific profile of preferences. One might ask whether, starting at a profile $P \in \mathcal{P}$ and arrangement (l,k) that is adequate at P, the comparative statics with respect to l and k are consistent. The following examples demonstrate that this is not so. It may be that, depending on P, increasing k renders a previously adequate arrangement inadequate, or the opposite. In other words, the effect of the increase to k is specific to P and l.

Example 1. *Either incrementing or decrementing l or k can render an adequate arrangement inadequate.*

Suppose $|D| = |H| = 3$ and consider $P \in \mathcal{P}$ such that for each $i = 1,2,3$,[10]

P_{h_i}	P_{d_i}
d_1	h_1
d_2	h_2
d_3	h_3
h_i	d_i

For P, $(2,2)$ is adequate: the interview matching is ν such that $\nu(h_1) = \nu(h_2) = \{d_1,d_2\}$ and $\nu(h_3) = \{d_3\}$. So, the (l,k)-matching is μ such that for each $i = 1,2,3$, $\mu(h_i) = d_i$, which is the unique stable matching.

We now observe that if we increment or decrement either l or k by one, the arrangement is no longer adequate for P. In other words, none of $(1,2)$, $(3,2)$,

[10] This can be embedded into a larger problem instance.

$(2,1)$, or $(2,3)$ are adequate for P. We summarize the interview matching and the (l,k)-matching for each of these below.

(l,k)	interview matching	(l,k)-matching
$(1,2)$	$\nu(h_1)=\nu(h_2)=\{d_1\},\nu(h_3)=\{d_2\}$	$\mu(h_1)=d_1,\mu(h_3)=d_2,\mu(h_2)=h_2,\mu(d_3)=d_3$
$(3,2)$	$\nu(h_1)=\nu(h_2)=D,\nu(h_3)=\{\}$	$\mu(h_1)=d_!,\mu(h_2)=d_2,\mu(h_3)=h_3,\mu(d_3)=d_3$
$(2,1)$	$\nu(h_1)=\{d_1,d_2\},\nu(h_2)=\{d_3\},\nu(h_3)=\{\}$	$\mu(h_1)=d_1,\mu(h_2)=d_3,\mu(h_3)=h_3,\mu(d_2)=d_2$
$(2,3)$	$\nu(h_1)=\nu(h_2)=\nu(h_3)=\{d_1,d_2\}$	$\mu(h_1)=d_1,\mu(h_2)=d_2,\mu(h_3)=h_3,\mu(d_3)=d_3$

All four of the (l,k)-matchings are unstable. ∘

The mechanics of Example 1 are robust and it is not by accident that $(2,2)$ is adequate to start with. The preferences in the example have a particularly salient configuration, which we focus on here. A profile $P\in\mathcal{P}$ has **common preferences** if all doctors rank the hospitals in the same way, and all hospitals rank the doctors in the same way. To further restrict the definition, we also require that each doctor finds each hospital acceptable and each hospital finds each doctor acceptable. That is, for each pair $d,d'\in D$ and each pair $h,h'\in H$, $P_d|_H=P_{d'}|_H$, $P_h|_D=P_{h'}|_D$, $d\,P_h\,h$, and $h\,P_d\,d$.[11]

As we see from Example 1, a result like Proposition 1 does not hold if we restrict ourselves to common preferences. Our next result is a characterization of homogeneous arrangements that are adequate for common preferences.[12]

Proposition 2. *Let $P\in\mathcal{P}$ be such that there are common preferences. A homogeneous arrangement (l,k) is adequate at P if and only if $l=k$ or $l,k\geq\min\{|D|,|H|\}$.*

Proof. Let $\{d_t\}_{t=1}^{|D|}$ and $\{h_t\}_{t=1}^{|H|}$ be enumerations of D and H, respectively, such that every hospital prefers d_t to d_{t+1} and every doctor prefers h_t to h_{t+1}. Let $m=$

[11] Under common preferences there is obviously a unique stable matching.

[12] The characterization of Proposition 2 does not hold for arrangements that are not homogeneous. For a counterexample, suppose $|D|=4$, $|H|=3$, there is $d\in D$ such that $\kappa_d=3$, for each $d'\in D\setminus\{d\},\kappa_{d'}=2$, there is $h\in H$ such that $\iota_h=4$, and for each $h'\in H\setminus\{h\},\iota_{h'}=2$. For any common preferences, (ι,κ) is adequate.

$\min\{|D|,|H|\}$. There is a unique stable matching μ^*, such that for each $t = 1,\dots,m$, $\mu^*(h_t) = d_t$.

Let ν be the interview matching under (l,k) and μ be the (l,k)-matching.

First, we show that (l,k) is adequate for P only if $l = k$ or $l,k \geq \min\{|D|,|H|\}$. Suppose $l \neq k$. If $l < k$ and $l < \min\{|D|,|H|\}$, then for each $t = 1,\dots,k$, $\nu(h_t) = \{d_1,\dots,d_l\}$. In particular, $d_k \notin \nu(h_k)$ so $\mu(h_k) \neq d_k$. On the other hand, if $l > k$ and $k < \min\{|D|,|H|\}$, then for each $t = 1,\dots,l$, $\nu(d_t) = \{h_1,\dots,h_k\}$. In particular, $h_l \notin \nu(d_l)$ so $\mu(d_l) \neq h_l$. In either case, the (l,k)-matching is not stable so (l,k) is not adequate.

Now, we show that if $l = k \leq m$, then (l,k) is adequate. For each $t = 1,\dots,m$, let $\underline{t} = \lfloor \frac{t-1}{l} \rfloor$. Then, for each $t = 1,\dots,i$, $\nu(h_t) = \{d_{\underline{t}+1},\dots,d_{\underline{t}+l}\}$ and $\nu(d_t) = \{d_{\underline{t}+1},\dots,d_{\underline{t}+l}\}$. Thus, for each $t = 1,\dots,m$, $\mu(h_t) = d_t$. So (l,k) is adequate at P.

Finally, if $l,k \geq m$, then for each $t = 1,\dots,m$, $\nu(d_t) \supseteq \{h_1,\dots,h_m\}$. Since $h_t \in \nu(d_t)$, $h_t = \mu(d_t)$. So (l,k) is adequate at P. □

If the hospitals' interview capacity is fixed at some specific l, an important policy decision involves where to set the doctors' interview cap, k. Proposition 2 says that the optimal value for k is exactly at l whether the objective is to minimize the number of blocking pairs or to maximize the match rate (the proportion of positions that are filled). Our next result sheds light on this objective.

Proposition 3. *Fix the hospitals' interview capacity at l and consider k and k' such that either $k' < k \leq l$ or $l \leq k < k'$. Suppose $P \in \mathcal{P}$ has common preferences. The (l,k')-matching has more blocking pairs and a weakly lower match rate than the (l,k)-matching.*

Proof. Let $P \in \mathcal{P}$ be such that there are common preferences. Let $\{d_t\}_{t=1}^{|D|}$ and $\{h_t\}_{t=1}^{|H|}$ be enumerations of D and H, respectively, such that every hospital prefers d_t to d_{t+1} and every doctor prefers h_t to h_{t+1}.

Let $m = \min\left\{\left\lfloor \frac{|H|}{k} \right\rfloor, \left\lfloor \frac{|D|}{l} \right\rfloor\right\}$. The interview matching is such that for each d_t, if $t \leq ml$,

$$\nu(d_t) = \{h_{(n-1)k+1},\dots,h_{nk}\} \text{ where } n \text{ is such that } (1-n)l < t \leq nl$$

if $ml < t \leq (m+1)l$,

$$\nu(d_t) = \begin{cases} \{h_{mk+1}, \ldots, h_n\} & \text{if } |H| \geq mk+1 \\ \emptyset & \text{otherwise} \end{cases} \quad \text{where } n = \min\{|H|, (m+1)k\}$$

and if $(m+1)n < t$, $\nu(d_t) = \emptyset$.

We first consider the case where when $k > l$ and show that the number of matched hospitals is decreasing in k and that the number of blocking pairs is increasing in k.

Given P and its restriction to ν, the (l,k)-matching, μ, at P is such that for each d_t, if $t \leq ml$,

$$\mu(d_t) = h_{(n-1)k+(t \bmod l)}, \text{ where } n \text{ is such that } (n-1)l < t \leq nl,$$

if $ml < t \leq (m+1)l$,

$$\mu(d_t) = \begin{cases} h_{mk+(t \bmod l)} & \text{if } |H| \geq mk + (t \bmod l) \\ d_t & \text{otherwise,} \end{cases}$$

and if $(m+1)l < t$, $\mu(d_t) = d_t$.

Let $n = \min\{|H| - mk, |D| - ml\}$. Given the (l,k)-matching above, the set of matched hospitals is

$$\{h_{ik+s} : i = 0, \ldots, m-1, s = 1 \ldots, l\} \cup \{h_t : t = mk+1, \ldots, mk+n\}.$$

Therefore, the number of matched hospitals is $ml + n$. Holding l fixed, this is decreasing in k.

The (l,k)-matching is blocked by all pairs consisting of an unmatched hospital and any doctor with a higher index. That is, $(h_t, d_{t'})$ such that $t \leq mk$, $t-1 \bmod k \geq l$ and $t' > t$. These are the only pairs that block it. Thus, the number of blocking pairs is

$$\sum_{n=0}^{m-1} \sum_{i=l+1}^{k} |D| - (nk+i).$$

Holding l fixed, this is increasing in k.

Now, we consider the case where $k < l$ and show that the number of matched hospitals is increasing in k and the number of blocking pairs is decreasing in k.

Given P and its restriction to ν, the (l,k)-matching at P is such that for each h_t, if $t \leq mk$,

$$\mu(h_t) = d_{(n-1)l+(t \bmod k)}, \text{ where } n \text{ is such that } (n-1)l < t \leq nl,$$

if $mk < t \leq (m+1)k$,

$$\mu(h_t) = \begin{cases} h_{ml+(t \bmod k)} & \text{if } |D| \geq ml + (t \bmod k) \\ h_t & \text{otherwise} \end{cases}$$

and if $(m+1)k < t$, $\mu(h_t) = h_t$.

Let $n = \min\{|H| - mk, |D| - ml\}$. Given (l,k)-matching above, the set of matched hospitals is $\{h_t : t \leq mk + n$. Therefore, the number of matched hospitals is $mk + n$. Since $k < l$, this is weakly increasing in k.

The (l,k)-matching is blocked by all pairs consisting of an unmatched doctor and any hospital with a higher index. That is, $(d_t, h_{t'})$ such that $t \leq ml$, $t-1 \bmod l \geq k$ and $t' > t$. These are the only pairs the block it. Thus, the number of blocking pairs is

$$\sum_{n=0}^{m-1} \sum_{i=k+1}^{l} |D| - (nl + i).$$

Holding l fixed, this is decreasing in k. □

5 Simulations

Our analytical results are of two sorts. On one hand, Theorem 1 applies without restrictions on preferences. However, it only has something to say about doctors' welfare and only in regards to perturbations to an equilibrium arrangement. On the other hand, when we focus on common preferences, Propositions 2 and 3 deliver a clearcut policy prescription. In this section, we use simulations to bridge the gap.

This allows us to consider how changes in the doctors' interview capacities affect hospitals' welfare, match rates, stability, and so on, in a more general setting.

While there is evidence that preferences do indeed have a common component (Agarwal, 2015; Rees-Jones, 2018), agents care about "fit" as well. Moreover, an idiosyncratic component is to be expected. We adopt the random utility model of Ashlagi et al. (2017).[13] Each hospital $h \in H$ has a common component to its quality, x_h^C, and a "fit" component, x_h^F. Similarly, each doctor $d \in D$ has a common component to her quality, x_d^C and a fit component, x_d^F. The utilities that h and d enjoy from being matched to one another are

$$u_h(d) = \beta x_d^C - \gamma \left(x_h^F - x_d^F\right)^2 + \varepsilon_{hd}$$

and

$$u_d(h) = \beta x_h^C - \gamma \left(x_h^F - x_d^F\right)^2 + \varepsilon_{dh},$$

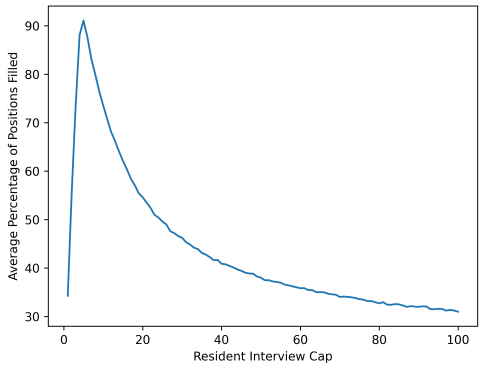
respectively, where ε_{hd} and ε_{dh} are drawn independently from the standard logistic distribution. Each x_h^C, x_h^F, x_d^C, and x_d^F is drawn independently from the uniform distribution over $[0,1]$. The coefficients β and γ weight the common and fit components, respectively. When β and γ are both zero, preferences are drawn uniformly at random. As $\beta \to \infty$, these approach common preferences. As γ increases, preferences become more "aligned": the fit, which is orthogonal to the common component, becomes more important.

Our simulated market has 400 hospitals.[14] We have set the number of doctors at 470.[15] The parameters for the random utility model are $\beta = 40$ and $\gamma = 20$. Since our interest is in the effects of changes to doctors' interview capacities,

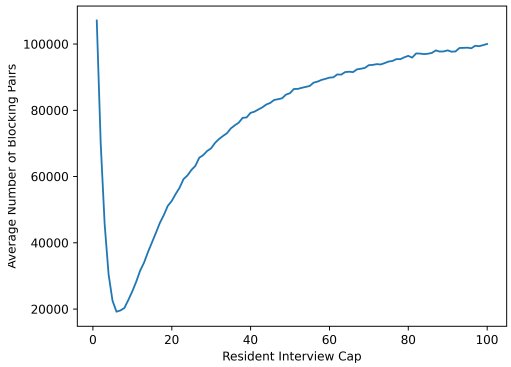
[13] This, in turn, is adapted from Hitsch et al. (2010).

[14] The NRMP match is broken down into smaller matches by specialty. In 2020, among 50 specialties for PGY-1 programs, the largest had 8,697 positions, the 10th largest had 849 positions, the 25th largest had 38 positions, the 49th largest had one position, and the smallest had no positions. This data is available from the NRMP. Our chosen number of hospitals is comparable to the 70th percentile among specialties (that is, 70% of specialities are smaller than this).

[15] There were, on average, 0.85 PGY-1 positions per applicant in the 2020. Our chosen number of doctors reflects this ratio.



(a) The average match rate is highest at $k = 5$ (90.655 pairs).



(b) The average number of blocking pairs is lowest at $k = 7$ (22,080 pairs).

Figure 1: We vary k from 1 to 100 with l fixed at 25.

we fix hospital interview capacities at $l = 25$. We discuss the robustness of our findings with regards to our choices of model and parameter values in Appendix E.

Our first simulation results involve varying k from 1 to 100.[16] Figure 1a shows that the match rate increases and then decreases. On the other hand, Figure 1b shows that the number of blocking pairs decreases and then increases. These results are consistent with Proposition 3. However, since preferences are not common, the match rate does not reach 100% and the number of blocking pairs remains positive even at the optimal k—that is, the arrangement is not adequate. Moreover, the optimal k does not equal l.

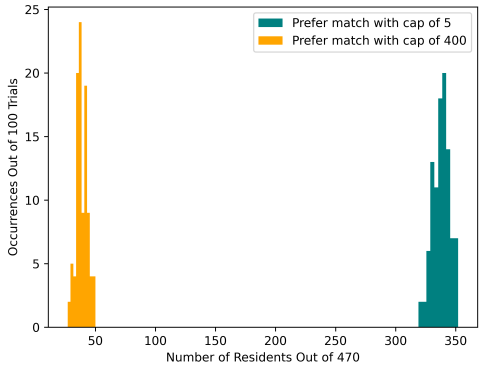
Our next set of results evaluate a hypothetical policy of restricting doctors to a maximum of 7 interviews. We choose this as a candidate policy since it is optimal in terms of minimizing the average number of blocking pairs (Figure 1b). We compare this policy with the benchmark of no intervention where doctors are completely unconstrained.

[16] We have chosen this upper bound to be high enough that further increases have little effect. Thus, we interpret this as doctors being essentially unconstrained in how many interviews they can accept.

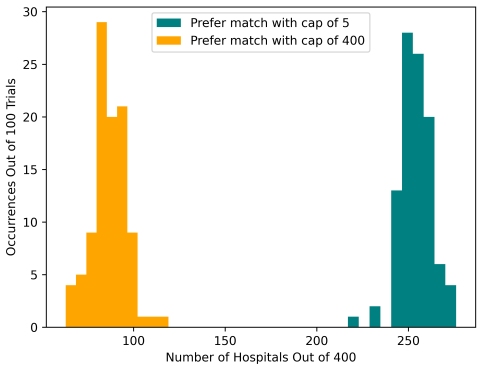
Figure 2a shows the distribution of the number of doctors who prefer their match under the optimal k over the benchmark as well as the distribution of those with the opposite preference. We see that the former is considerably higher than the latter. Though Theorem 1 applies only when the starting arrangement is adequate, Figure 2a shows that the lesson from that result does extend beyond. Figure 2b shows the same distributions, except for hospitals. Despite the fact that Theorem 1 does not address hospitals' welfare, our simulations show that more hospitals prefer the $k = 7$ than leaving the doctors unconstrained. The policy also has the benefit of drastically decreasing the number of blocking pairs. Figure 2c shows the distribution of excess blocking pairs when we compare the benchmark matching where doctors are unconstrained to the matching under $k = 7$. Finally, we compare the distribution of interviews among the doctors between the two arrangements in Figure 2d. The constraint limiting doctors to $k = 7$ interviews binds for many doctors. One implication is that significantly more doctors receive zero interviews when they are unconstrained. This is consistent with the intuition that if interviews were costless for doctors, then highly sought after doctors would hoard interviews and others would be left with nothing.

Finally, we consider the possibility that the NRMP could set not only a cap on interviews that doctors can accept, but can also control the number of interviews that hospitals offer. From Proposition 2, we know that if preferences are common, the match rate would be maximized where $l = k$. Figure 3 shows that, even when preferences are not exactly common, there are still optimal combinations of l and k, but they typically involve $l > k$.

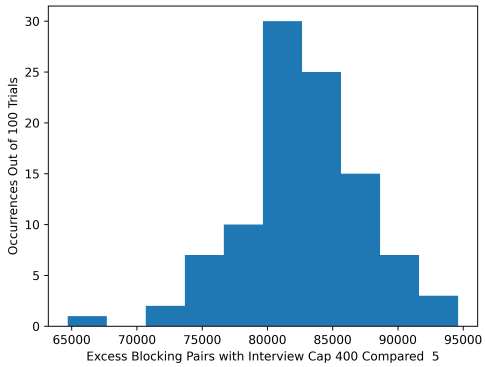
In Appendix D, we modify our model to consist of two tiers of doctors and contrast simulations under naïve choices by hospitals (as above) to those under a heuristic for hospitals where every choice involves a "safety" doctor from the lower tier. The only qualitative difference is in the match rate: rather than being unmatched, hospitals match with safety candidates. The welfare based comparisons are very similar.



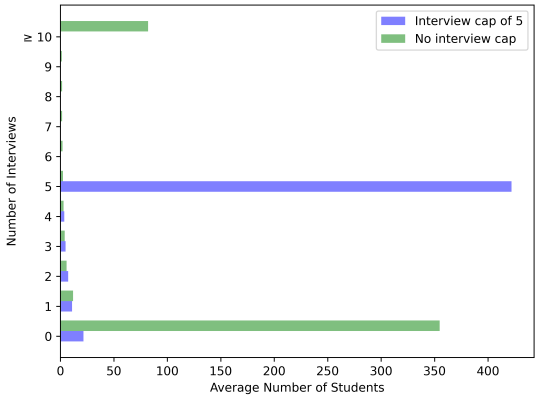
(a) Distribution of the number of doctors who prefer their match at $k = 7$ over being unconstrained and vice versa.



(b) Distribution of the number of hospitals that prefer their match at $k = 7$ over the doctors being unconstrained and vice versa.



(c) Distribution of the number of excess blocking pairs when doctors are unconstrained over the number of such pairs at $k = 7$.



(d) Distributions of interviews at $k = 7$ and without a cap. The uncapped distribution vanishes with the number of interviews reaching the hundreds.

Figure 2: Comparisons of the intervention of capping doctors' interview capacities at $k = 7$ to leaving doctors unconstrained.

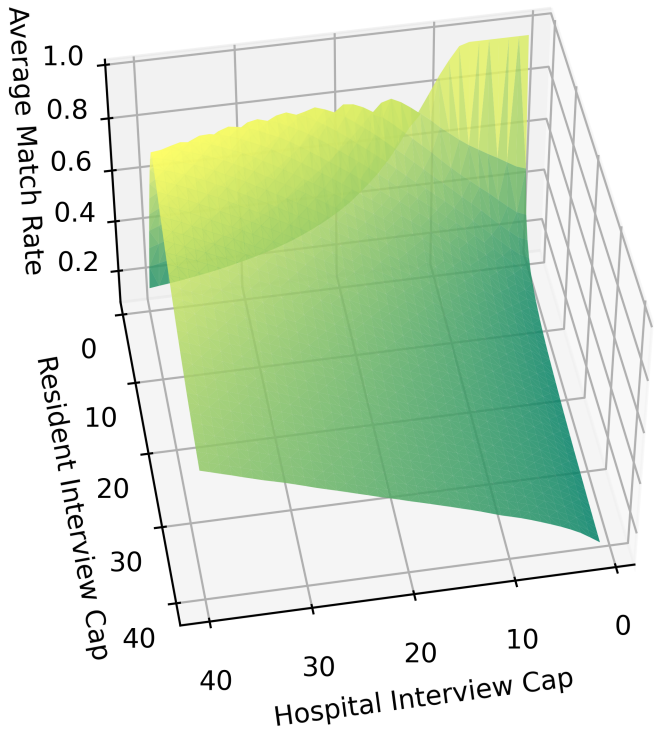


Figure 3: Match rate as a function of l and k

6 Conclusion

The 2020-21 global pandemic has had a significant impact on the way interviews are conducted. It has also impacted the distribution of interviews among doctors. Through our theoretical results and simulations, we argue that the 2021 NRMP Match was likely inferior to previous years.

In future years, the NRMP should consider policies to mitigate these effects. Our analysis supports the idea of interview caps and our simulations provide evidence that such a policy would reduce the bottleneck created by the interview phase. Such caps can be implemented with very limited centralization, for instance, using a ticket system.

Even if such interventions are not possible in the very short run, our policy prescription is that residency programs should be advised to increase the number of candidates they interview relative to previous years.

Design of a fully centralized clearinghouse is an area that remains open. As earlier work on the interview pre-markets have shown, strategic analysis has only been tractable under very stringent assumptions (Kadam, 2021; Lee and Schwarz, 2017; Beyhaghi, 2019). Nonetheless, the current paper adds to the evidence (along with Echenique et al. (2020)) that a more holistic approach that includes the interview stage is critical.

This interview-driven bottleneck is likely a factor in other matching contexts as well, including fully decentralized labor markets. For example, we expect it will affect the junior market for economists. This labor market typically consists of short interviews followed by on-campus visits. Physical constraints typically limit both the number of short interviews and on-campus visits a candidate is able to accept. With virtual interviews and virtual "fly-outs," we expect candidates to accept more of both than they would otherwise. As a result, we could expect the same bottleneck in the economics job market as in the NRMP match.

References

Agarwal, Nikhil (2015) "An Empirical Model of the Medical Match," *American Economic Review*, Vol. 105, No. 7, pp. 1939–78, July. 24

Ashlagi, Itai, Yash Kanoria, and Jacob D. Leshno (2017) "Unbalanced Random Matching Markets: The Stark Effect of Competition," *Journal of Political Economy*, Vol. 125, No. 1, pp. 69–98. 24, 42, 44, 45

Beyhaghi, Hedyeh (2019) "Approximately-optimal Mechanisms in Auction Design, Search Theory, and Matching Markets," Ph.D. dissertation, Cornell University. Chapter 5: Two-Sided Matching with Limited Number of Interviews. 5, 29

Chade, Hector and Lones Smith (2006) "Simultaneous Search," *Econometrica*, Vol. 74, No. 5, pp. 1293–1307. 6, 38

Echenique, Federico, Ruy Gonzalez, Alistair Wilson, and Leeat Yariv (2020) "Top of the Batch: Interviews and the Match." 2, 3, 5, 6, 9, 29, 34, 42

Gale, David and Lloyd Shapley (1962) "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, Vol. 69, pp. 9–15. 2

Hitsch, Gunter J., Ali Hortaçsu, and Dan Ariely (2010) "Matching and Sorting in Online Dating," *American Economic Review*, Vol. 100, No. 1, pp. 130–63, March. 24

Kadam, Sangram V (2021) "Interviewing in Matching Markets with Virtual Interviews." 5, 29

Lee, SangMok (2016) "Incentive Compatibility of Large Centralized Matching Markets," *The Review of Economic Studies*, Vol. 84, No. 1, pp. 444–463, 09. 42, 44, 45, 46

Lee, Robin S. and Michael Schwarz (2017) "Interviewing in two-sided matching markets," *The RAND Journal of Economics*, Vol. 48, No. 3, pp. 835–855. 5, 29

Rees-Jones, Alex (2018) “Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match,” *Games and Economic Behavior*, Vol. 108, pp. 317–330. 24

Roth, Alvin E. and Elliott Peranson (1999) “The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design,” *American Economic Review*, Vol. 89, No. 4, pp. 748–780, September. 2, 5

Roth, Alvin E. and Marilda A. Oliveira Sotomayor (1990) *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs: Cambridge University Press. 8, 11

Appendices

A Role of Interviews: Preference Formation and Coordination

The interviews that precede the NRMP match serve at least two important purposes. The more obvious one is preference formation: while agents have a prior sense of their preferences, they update their preferences based on information revealed by the interviews. Less obvious is the coordination of whom to rank. A pair of agents can only be matched if they *rank one another*. In a market with hundreds, if not thousands, of potential partners, even formulating—much less submitting—a ranking of all possible partners is impractical. Thus agents submit only a portion of their preferences. Interviews function as a device to coordinate which portions they submit.

The second role of interviews that we have mentioned above is important even in a world where interviews are completely uninformative.

A natural question is to ask how the analysis of this paper is affected when we take preference formation into account. We first observe that we have drawn conclusions about doctors' welfare (Theorem 1) and stability (Proposition 2 and Proposition 3). These comparisons are based on exogenous preferences. If preference formation is part of the model, then preferences are endogenous and there is no basis for such comparisons. However, we explain below how even here the match rate, which is an objective measure, is maximized when there is balance between the two sides' interview capacities.

Represent the preferences of the agents prior to the interviews by the **pre-interview preferences,** $\overline{\mathbf{P}} \in \mathcal{P}$. As with Propositions 2 and 3, we assume that $\overline{P}$ has common preferences.

Given an interview matching ν, for each agent $i \in D \cup H$, let $\mathbf{P_i^\nu}$ be i's **post-interview preferences**. These are the preferences that i forms through the interview process. The only requirement for P_i^ν is that $\nu(i)$ is the set of partners that i ranks as acceptable. Note that we do not make any other assumptions about how P^ν relates to $\overline{P}$. In fact, we leave the exact orderings unspecified as this is unnecessary to draw conclusions about the match rate.

Since the pre-interview preferences are common, there is a unique interview matching, ν, that is stable with respect to these preferences. Given ν, unlike the two-phase process in Section 2, the next step is to use the *post-interview preferences*, $\mathcal{P}^\nu$ (as opposed to $(\overline{P}_i|_{\nu(i)})_{i \in D \cup H}$) as the input to the doctor-proposing DA algorithm. The output is what we call the **$(\mathbf{l},\mathbf{k})$-matching with updating**.

Note that if the interviews either lead to a matching ($l = k = 1$) or do not impose any constraints ($l, k \geq \min\{|D|,|H|\}$), then a total of $\min\{|D|,|H|\}$ pairs form. In the proof of below, we demonstrate that when $l = k$, the (l,k)-matching with updating matches as many pairs as this "first best" benchmark. Increasing the gap between l and k causes the (l,k)-matching with updating to fall short of this benchmark regardless of what the updated preferences are. This is a straightforward consequence of the pigeon hole principle.

Proposition 4. *Fix the hospitals' interview capacity at l and consider k and k'*

such that either $k' < k \leq l$ *or* $l \leq k < k'$*. The* (l,k')*-matching with updating has a weakly lower match rate than the* (l,k)*-matching with updating.*

Proof. Let $\{d_t\}_{t=1}^{|D|}$ and $\{h_t\}_{t=1}^{|H|}$ be enumerations of D and H, respectively, such that every hospital prefers d_t to d_{t+1} and every doctor prefers h_t to h_{t+1} under $\overline{P}$.

Let $m = \min\left\{\left\lfloor \frac{|H|}{k} \right\rfloor, \left\lfloor \frac{|D|}{l} \right\rfloor\right\}$. For each $n = 1,\dots,m$, let

$$D_n = \{d_t : (n-1)l < t \leq nl\}$$
and
$$H_n = \{h_t : (n-1)k < t \leq nk\}.$$

By definition of m, either $|D| < (m+1)l$ or $|H| < (m+1)k$. Let $t_D = \min\{|D|, (m+1)l\}$, $t_H = \min\{|H|, (m+1)k\}$,

$$D_{m+1} = \{d_t : ml < t \leq t_D\},$$
and
$$H_{m+1} = \{h_t : mk < t \leq t_H\}$$

The unique pairwise stable interview matching, ν, is such that for each $n = 1,\dots,m+1$, every doctor in D_n is interviewed by every hospital in H_n—that is, for each $d \in D_n$, $\nu(d) = H_n$ and for each $h \in H_n, \nu(h) = D_n$. Moreover, every doctor or hospital not in one of these sets has no interviews—that is, if there is d_t such that $t > t_D$, then $\nu(d_t) = \emptyset$ and if there is h_t such that $t > t_H$, then $\nu(h_t) = \emptyset$.

Now we consider the (l,k)-matching with updating, μ. Having received no interviews, any doctor or hospital with index higher than t_D or t_H, respectively, is necessarily unmatched regardless of the post-interview preferences. Since each agent's post-interview preferences only rank those agents in one's interview matching as acceptable, for each $n = 1,\dots,m+1$, we have that

1. for each $d \in D_n$, $\mu(d) \in H_n \cup \{d\}$ and
2. for each $h \in H_n$, $\mu(h) \in D_n \cup \{h\}$.

If $l < k$, then for each $n = 1, \ldots, m$, $|D_n| = l < k = |H_n|$. So exactly $k - l$ hospitals in H_n are unmatched. There are $m(k-l)$ such hospitals. If there are fewer hospitals in H_{m+1} than doctors in D_{m+1}, then $|H_{m+1}| - |D_{m+1}|$ additional hospitals are unmatched. That is, the number of unmatched hospitals among those in H_{m+1} is $\max\{0, |H_{m+1}| - |D_{m+1}|\} = \max\{0, (t_H - mk) - (t_D - m_l)\} = \max\{0, (t_H - t_D) - m(k-l)\}$. So the total number of unmatched hospitals is

$$m(k-l) + \max\{0, (t_H - t_D) - m(k-l)\},$$

which is weakly increasing in k.

If $k < l$, then every hospital h_t such that $t \leq mk$ is matched by μ. As when $l < k$, $\max\{0, (t_H - t_D) - m(k-l)\}$ hospitals among those in H_{m+1} are unmatched by μ. Every hospital h_t such that $t > t_H$ is also unmatched. Thus, the number of hospitals that μ leaves unmatched is $\max\{0, |H| - (mk + |D| - ml)\} = \max\{0, |H| - |D| + m(l-k)\}$, which is weakly decreasing in k. □

Proposition 4 shows that the bottleneck of imbalanced interview capacities occurs at the interview stage. In this sense, the preference formation role of interviews is orthogonal to our main point regarding the harm such imbalance causes.

B Doctor-optimal Interview Matching

The only difference between our model and that of Echenique et al. (2020) is that we suppose that the interview matching is hospital-optimal rather than doctor-optimal. Their modeling choice is natural for the question they ask as it gives each doctor her *best* stable set of interviews. Thus, their result that most doctors match with hospitals they rank highly can only be stronger for other interview matchings. For our analysis, the doctor-optimal interview matching does not have this natural appeal. To the contrary, hospital-proposing DA is a reasonable approximation of the interview assignment process. Nonetheless, our results are not driven by this choice. The only proof that relies on this choice is that of Theorem 1. In this section, we show that the result holds even for the doctor-optimal interview

matching followed by the doctor-optimal final matching. In what follows, we use the same terminology and notation as before, with the understanding that the interview matching is doctor-optimal.

As in the statement of the theorem, suppose that for each $d \in D$, $\kappa_d \leq \kappa'_d$. We show below that the Lemmas 2 and 3 hold even with the change from the hospital-optimal interview matching to the doctor-optimal interview matching. The key is to establish that, in the interview phase, if a doctor d is rejected by a hospital h under capacities κ, then h rejects her under κ' as well. Given capacities $\bar{\kappa}$, let $A_d(m;\bar{\kappa})$ be the hospitals that d proposes to and $R_d(m;\bar{\kappa})$ be the hospitals that reject doctor d by the end of round m of the interview phase. We show that these sets are monotonic in $\overline{\kappa}$.

Claim 3. *For any positive integer* m*,*

$$R_d(m;\kappa) \subseteq R_d(m;\kappa')$$
$$A_d(m;\kappa) \subseteq A_d(m;\kappa')$$

Proof. We proceed by induction on m, the base case being $m = 1$. In the first round of the interview phase, each $d \in D$ proposes to her favorite hospitals up to her interview capacity. Since every doctor proposes to at least as many hospitals under κ' as under κ, every hospital receives at least as many proposals under κ' as under κ. Therefore, if a doctor d is rejected by a hospital h in the first round of the interview phase under κ, she is also rejected by h in the first round under κ'. Now consider a round $m > 1$ of the interview phase and suppose for each doctor d that $R_d(m-1;\kappa) \subseteq R_d(m-1;\kappa')$ and $A_d(m-1;\kappa) \subseteq A_d(m-1;\kappa')$. In round m, each $d \in D$ proposes to her favorite hospitals that have not yet rejected her up to her interviewing capacity. Under κ, d proposes to her κ_d favorite hospitals in $H \setminus R_d(m-1;\kappa)$. Under κ', she proposes to her κ'_d favorite hospitals in $H \setminus R_d(m-1;\kappa')$. By the inductive hypothesis, $H \setminus R_d(m-1;\kappa') \subseteq H \setminus R_d(m-1;\kappa)$. Therefore, if d proposes to h under κ, either she proposes to h under κ' as well (she is choosing more hospitals from a smaller set of options) or she has already proposed to and has been rejected by h under κ'. In either case, if $h \in A_d(m;\kappa)$, then $h \in A_d(m;\kappa')$.

Since each hospital h receives more proposals but its capacity does not change, if h rejects doctor d under κ, she also rejects doctor d when choosing from a larger set of doctors who have proposed to it under κ'. Therefore, if $h \in R_d(m;\kappa)$, then $h \in R_d(m;\kappa')$. □

We now explain how Claim 3 implies that Lemmas 2 and 3 hold even when we switch to the doctor-optimal interview matching. Let ν and μ be the interview and final matchings respectively, under (ι,κ). Similarly, let ν' and μ' be the interview and final matchings under (ι,κ').

Lemma 2 says that for each $d \in D$, if $h \in \nu'(d) \setminus \nu(d)$, then $\mu(d)\ P_d\ h$. Given $\bar{\kappa}$, let $R_d(\bar{\kappa})$ denote the set of hospitals that reject d in any round of the interview phase under capacities $\bar{\kappa}$. By Claim 3, $R_d(\kappa) \subseteq R_d(\kappa')$. Under κ, $\nu(d)$ consists of d's κ_d most preferred hospitals in $H \setminus R_d(\kappa)$. That is, the κ_d highest-ranked hospitals that did not reject her. Under κ', $\nu'(d)$ comprises d's κ'_d most preferred hospitals in $H \setminus R_d(\kappa')$. As $H \setminus R_d(\kappa') \subseteq H \setminus R_d(\kappa)$, if $h' \in \nu'(d) \setminus \nu(d)$, then for every $h \in \nu(d)$, $h\ P_d\ h'$. In words, since d is interviewed by h under κ, she was not rejected by h under κ. As more hospitals rejected d under κ' than under κ, h does not reject d under κ. Therefore, d could have proposed to h under κ, but she chose not to. Therefore, by revealed preference, she prefers all hospitals in $\nu(d)$ to any of her "new" interviews under κ' (those in $\nu'(h) \setminus \nu(h)$).

Lemma 3 said that if $d \in \nu'(h)$, $d' \in \nu(h)$, and $d'\ P_h\ d$, then $d' \in \nu'(h)$. By Claim 3, d' proposes to at least as many hospitals in the interview phase under κ' as under κ. Since d' proposes to h under κ, she also proposes to h under κ'. Each hospital h accepts its ι_h favorite applicants, so if it accepts d, it must also accept d'.

Since Lemmas 2 and 3 hold, the remainder of the proof follows exactly as in Section 3.

C Choice Functions for Interview Phase

In the interview phase, we compute a many-to-many matching. However, each doctor and each hospital only ultimately matches to at most one other partner, and each has strict preferences over partners. This necessitates the definition of a choice function over sets of partners. Consistent with the assumption of non-strategic behavior with complete information, we focus on acceptant choice functions that are responsive to preferences over partners and constrained by interview capacity. That is, given $P \in \mathcal{P}$,

- From the set $H' \subseteq H$, each $d \in D$ chooses the κ_d best elements of H' according to P_d:

$$C_d(H') = \begin{cases} \{h \in H' : h \, P_d \, d\} & \text{if } |\{h \in H' : h \, P_d \, d\}| \leq \kappa_d \text{ and} \\ B \subseteq \{h \in H' : h \, P_d \, d\} & \text{such that } |B| = \kappa_d \text{ and for each } h \in B \\ & \text{and each } h' \in H' \setminus B, h \, P_d \, h' \text{ otherwise.} \end{cases}$$

- From the set $D' \subseteq D$, each $h \in H$ chooses the κ_h best elements of D' according to P_h:

$$C_h(D') = \begin{cases} \{d \in D' : d \, P_h \, h\} & \text{if } |\{d \in D' : d \, P_h \, h\}| \leq \iota_h \text{ and} \\ B \subseteq \{d \in D' : d \, P_h \, h\} & \text{such that } |B| = \iota_h \text{ and for each } d \in B \\ & \text{and each } d' \in D' \setminus B, d \, P_h \, d' \text{ otherwise.} \end{cases}$$

D Epilogue: What Actually Happened

The 2021 "Match Day"—the day the NRMP announces the results of the match—was on March 19. In the words of the NRMP President and CEO, Donna L. Lamb,

> The application and recruitment cycle was upended as a result of the pandemic, yet the results of the Match continue to demonstrate strong

and consistent outcomes for participants.[17]

Indeed, contrary to our results, there was a 2.6 percent increase in PGY-1 positions filled.

In this section, we argue that there is reason to be skeptical about the claim that the 2021 match is "strong and consistent." In particular, we contend that focusing on match rates leads one to miss an effect that is analogous to our results.

We have assumed that hospitals naïvely extend interview invitations to their most preferred doctors first and that these trickle down to less preferred doctors. However, hospitals are not naïve in practice and use heuristics.[18] Most heuristics include an option that has a high match probability even if it is ranked relatively low.[19] We call these candidates "safety" candidates. When hospitals offer interviews in this way, an increase to k increases the proportion of hospitals that match with their safety candidates rather than the number of unfilled positions. In other words, hospitals tend to match with lower ranked doctors, the likelihood of matching for higher tier candidates decreases, and the likelihood of matching for lower tier candidates increases.

The very limited information that the NRMP has published so far supports this conjecture. We do not have access to the actual preferences of residency programs, nor their submitted rankings. Yet, the reported interview and ranking patterns in the 2020 NRMP Program Director Survey reveal a systematic preference for MD Seniors (graduating students from US MD medical schools) over DO Seniors (graduating from US DO medical schools) as well as MD and DO Grads (those who have previously graduated from US medical schools but had

[17] NRMP Press Release dated March 19, 2021.

[18] Optimal strategies depend critically on the probability of matching with a doctor conditional on interviewing her (Chade and Smith, 2006). In our context, these probabilities are dependent not only on other hospitals' preferences and strategies but also on the rest of the hospital's own interview choices. It is implausible that hospitals have this information and that they then compute the optimal portfolio of doctors to interview.

[19] Even the optimal solution has this property when being unmatched is very unattractive relative to matching with lower ranked candidates (Chade and Smith, 2006).

not matched). Nonetheless, the proportion of positions filled by MD Seniors declined slightly while the proportions of positions filled by the latter three increased slightly.[20] This lends support to our hypothesis that more positions were filled by safety candidates.

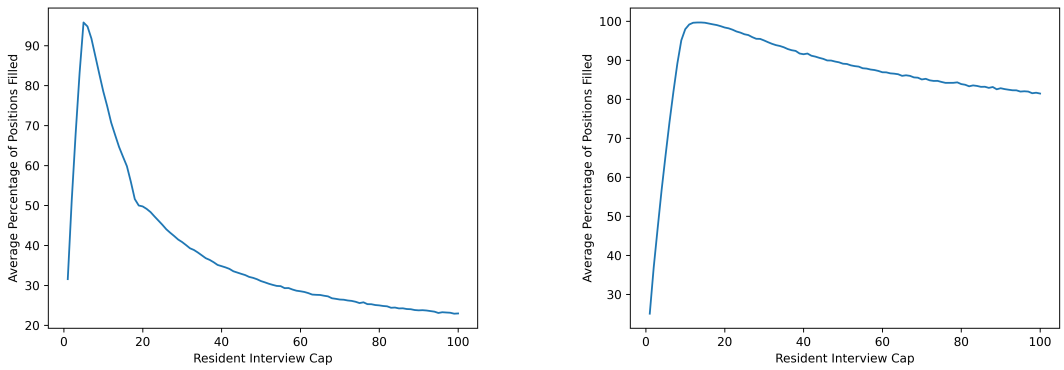
In what follows, we consider a variation of our model and contrast the matching patterns generated by the naïve behavior and a particular heuristic. The goal is to demonstrate the above explanation of how increased matches to safety candidates may mask the effects of increased interview capacities.

The change to the model is that we separate the doctors into two tiers—D_1 and D_2 (so $D_1 \cap D_2 = \emptyset$ and $D_1 \cup D_2 = D$)—such that every hospital prefers every doctor in D_1 to every doctor in D_2. We achieve this change by adding $\beta + \gamma + 10$ to $u_h(d)$, as specified in Section 5, for each $d \in D_1$. We have chosen D_1 such that $|D_1| = 200$.

To model behavior under the heuristic, we use the following choice functions for hospitals, as opposed to those defined in Appendix C. From a set $D' \subseteq D$, hospital $h \in H$ chooses a safety candidate from the second tier, if one is available, and fills the remaining slots by choosing the best options in D'. Note that in order to ensure that the interview matching (the hospital-proposing Deferred Acceptance) algorithm terminates, hospitals do not reconsider tentatively accepted interview offers. Thus, the choice function is parameterized by the number of interview offers to be made—in each round of the algorithm, the number of doctors a hospital chooses is its interview capacity minus the number of tentatively accepted offers in the previous round. Thus, when the hospital's preferences over the doctors are represented by P_h it chooses n doctors from $D' \subset D$ as follows: If $D \subseteq D_1$,

$$C_h(n, D') = \begin{cases} \{d \in D' : d \, P_h \, h\} & \text{if } |\{d \in D' : d \, P_h \, h\}| \leq n \text{ and} \\ B \subseteq \{d \in D' : d \, P_h \, h\} & \text{such that } |B| = n \text{ and for each } d \in B \\ & \text{and each } d' \in D' \setminus B, d \, P_h \, d' \text{ otherwise.} \end{cases}$$

[20] See the NRMP Reports for 2020 and 2021.



(a) Naïve choices

(b) Heuristic choices

Figure 4: We vary k from 1 to 100 with l fixed at 25 and display the corresponding match rate for naïve behavior by the hospitals, as well as the heuristic where each choice includes a safety candidate.

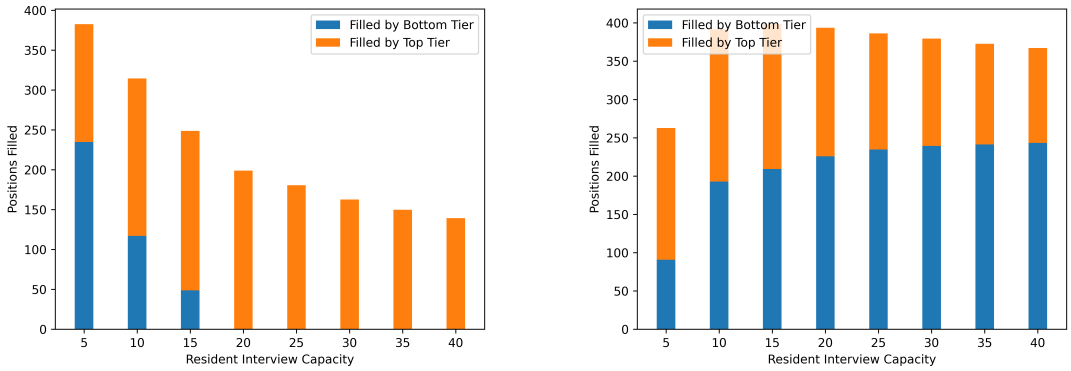
Otherwise, let d^s be chosen uniformly at random from $D' \cap D_2$.[21] Then,

$$C_h(n, D') = \begin{cases} \{d^s\} \cup \{d \in D' : d \, P_h \, h\} & \text{if } |\{d \in D' \setminus \{d^s\} : d \, P_h \, h\}| \leq n-1 \text{ and} \\ \{d^S\} \cup B & \text{where } B \subseteq \{d \in D' \setminus \{d^s\} : d \, P_h \, h\} \\ & \text{is such that } |B| = n-1 \text{ and for each } d \in B \\ & \text{and each } d' \in D' \setminus (B \cup \{d^s\}), d \, P_h \, d' \\ & \text{otherwise.} \end{cases}$$

We first observe that under the heuristic, the effect that increasing k has on the match rate all but disappears. In Figures 4a and 4b we show the effects for the naïve behavior and the heuristic respectively.[22] The dramatic difference in the match rates is entirely accounted for by the number of positions that are filled by second tier candidates. In other words, matching with lower ranked candidates is the alternative to being unmatched. This is easily seen in the comparison between Figures 5a and 5b.

[21] Hospitals randomize their safety choice d^s to avoid overlap.

[22] All parameter values are the same as in Section 5.



(a) Naïve choices

(b) Heuristic choices

Figure 5: For a range of values of k, with l fixed at 25, we see that under naïve choices more positions are unfilled as k increases. Under the heuristic, the match rate does not drop substantially, but positions are filled by lower ranked candidates. In either case, the number of positions filled by top tier candidates decreases with k, which is consistent with our theory.

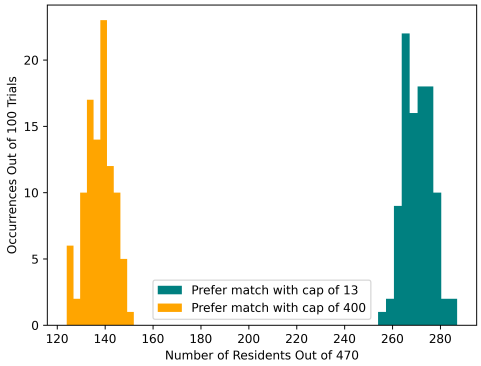
Other than the match rate, the remaining patterns that we presented in Section 5 persist with our modified model under the heuristic for hospitals' choices. Under the heuristic, lower ranked doctors who would be unmatched at lower k gain from the misallocation of interviews among higher ranked doctors. Nonetheless, more doctors prefer the outcome under a cap of 13 than prefer the benchmark with no cap as shown in Figure 6a.[23] As noted above, hospitals tend to match with lower ranked doctors and therefore tend to be worse off, as shown in Figure 6b. The result, as before, is driven by interview hoarding, which can be seen from the comparison of interview distributions with a cap of 13 and with no cap in Figure 6c.

E Robustness of Simulations: Choice of Random Utility Model and Parameters

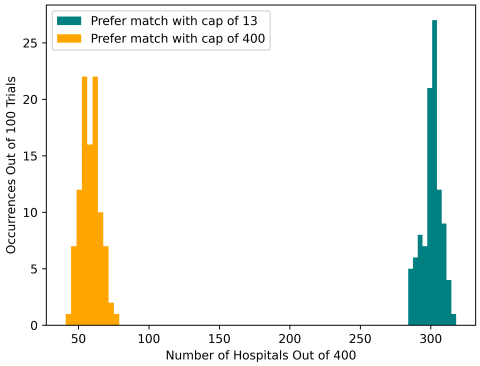
In Section 5, we have adopted the random utility model of Ashlagi et al. (2017). Moreover, we have presented our results for fixed values of several parameters: $\beta = 40$, $\gamma = 20$, $l = 25$, $|D| = 470$, and $|H| = 400$. In this appendix, we discuss the robustness of our findings with regards to these choices.

We start with the random utility model. A very natural alternative is that of Lee (2016), which is the model that Echenique et al. (2020) have adopted. This model does not include a fit component, so a doctor's utility from being matched to a hospital is a convex combination of a common (across all doctors) value and an idiosyncratic (to that doctor) value. A hospital's utility from being matched to a doctor is similarly comprised. Thus, each hospital $h \in H$ has a common component to its quality, x_h^C, and each doctor has a common component to her quality, x_d^C. Aside from this, for each doctor-hospital pair, d and h, ε_{dh} is the idiosyncratic value that d assigns to h and ε_{hd} is the idiosyncratic value that h assigns to d. Then, the utilities that h and d enjoy from being matched to one

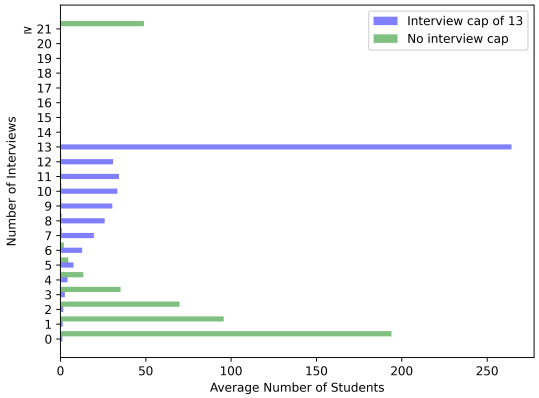
[23] We chose 13 as this maximizes the match rate for the tiered model under the heuristic.



(a) Distribution of the number of doctors who prefer their match at $k = 13$ over being unconstrained and vice versa.



(b) Distribution of the number of hospitals that prefer their match at $k = 13$ over the doctors being unconstrained and vice versa.



(c) Distributions of interviews at $k = 13$ and without a cap. The uncapped distribution vanishes when the number of interviews reach the hundreds.

Figure 6: Results for the tiered model with heuristic choices for the hospitals that parallel our results from Section 5.

another are

$$u_h(d) = \alpha x_d^C + (1-\alpha)\varepsilon_{hd}$$

and

$$u_d(h) = \alpha x_h^C + (1-\alpha)\varepsilon_{dh},$$

respectively.

The random variables x_h^C, x_d^C, ε_{dh}, and ε_{hd} are all independently drawn from the uniform distribution over $[0,1]$. The coefficients α and $(1-\alpha)$ weight the common and idiosyncratic components, respectively.

To make an apples-to-apples comparison between the two models, we set $\gamma = 0$ since the Lee (2016) model does not have a fit component. The remaining difference is the distribution of the idiosyncratic components: the standard logistic distribution in the former and the uniform distribution over $[0,1]$ in the latter. Though these distributions have different supports, we can relate the two models by considering, for each pair $d, d' \in D$ and each $h \in H$, the degree of correlation between $u_d(h)$ and $u_{d'}(h)$ as given by the Pearson correlation coefficient.[24] For the Lee (2016) model with parameter α, it is $\alpha^2/\alpha^2+(1-\alpha)^2$. For the Ashlagi et al. (2017) model with parameter β (and $\gamma = 0$), it is $\beta^2/\beta^2+(2\pi)^2$. Thus, utilities have the same linear correlation in the two models when

$$\alpha = \frac{\beta}{\beta + 2\pi}.$$

We display this relationship in Figure 7. The value of 40 that we have chosen for β in Section 5 corresponds to an α of 0.864.

Now consider γ. The higher γ is, the more aligned preferences are *across* the two sides of the market. However, when we compare two doctors, when $\gamma > 0$, the further apart they are in terms of their fit component, the less correlated their utilities are. Thus, since $\beta^2/\beta^2+(2\pi)^2$ is the correlation between the utilities of two doctors with the same fit component, it is an upper bound on the correlation between the preferences of any two doctors. Our chosen value of γ is on the same order of magnitude as β.

[24] Given the symmetry of both models, we could equivalently state this as the correlation between $u_h(d)$ and $u_{h'}(d)$ for for each pair $h, h' \in H$ and each $d \in D$.

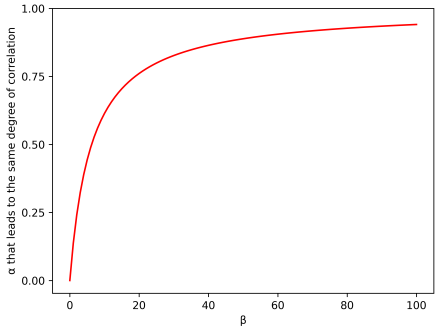


Figure 7: Locus of α and β pairs that result in the same linear correlation between two doctors' (hospitals') utilities from the same hospital (doctor) in the models of Lee (2016) and Ashlagi et al. (2017), respectively

.

We intend for our simulation results to be suggestive of how the insights from our analytical results may extend beyond the assumptions that we make for the sake of tractability. To this end, none of the choices that we have made are critical in driving the effects we describe in Section 5. We consider each of these choices in turn and make our point by focusing on how the match rate responds to different values of k as shown in Figure 1a. In other words, we show how this relationship changes as we vary our model and parameter choices. In what follows, we vary only one choice at a time, leaving fixed the other parameters as in Section 5.

In Figure 8, we consider various values of α and use the Lee (2016) model. As long as there is at least moderate correlation in preferences, we observe the effect, albeit with varying magnitude. Indeed, as the correlation between utilities grows, the effect becomes stronger, eventually converging to the one described by Proposition 3 when $\alpha = 1$.

In Figure 9, we consider various combinations of β and γ. Again, as β grows, the correlation in preferences increases and the effect becomes stronger. However, as discussed above, $\gamma > 0$ has the effect of decreasing the correlation between utilities of agents on the same side on average. Consistent with this under-

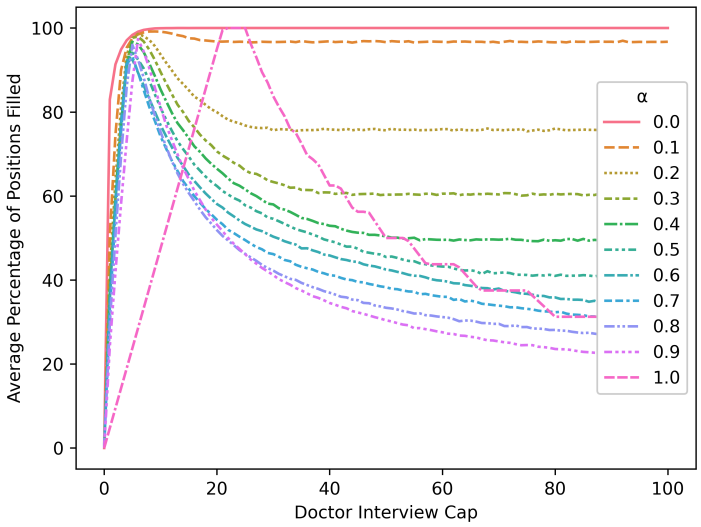


Figure 8: The average match rates for each k under the random utility model of Lee (2016) at various values of α.

standing, we see that for a fixed value of β, the lower γ is, the stronger the effect.

The next parameter we vary is the number of doctors. Recall that we fixed the number of hospitals at 400 for the simulations reported in Section 5. We leave this number fixed in Figure 10 but vary the number of doctors. For the cases where there are fewer than 400 doctors, the match rate is necessarily lower than 100%.

Keeping the ratio of hospitals to doctors the same, we next proportionally vary the size of the market in Figure 11. The value of k that maximizes the match rate decreases as the market becomes larger.

Finally, we consider the value of l that we have fixed at 25 for the simulations in Section 5. Though the effect persists no matter the value of l, the magnitude decreases as l increases as we show in Figure 12. This suggests that a policy that induces hospitals to interview more candidates would alleviate the problem to some extent.

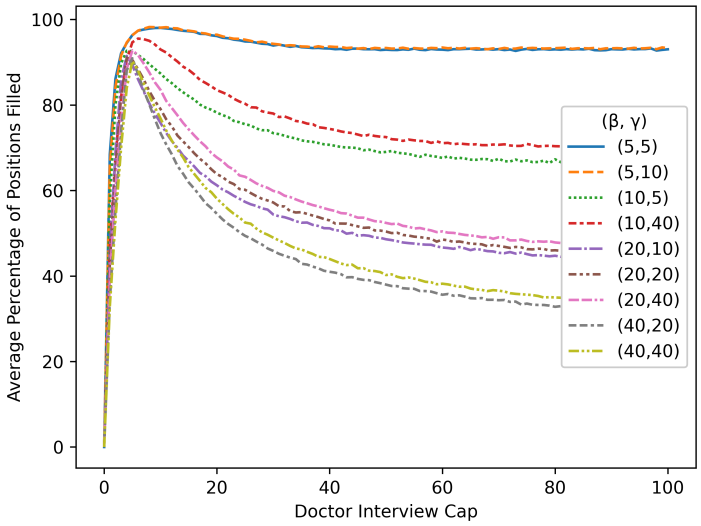


Figure 9: The average match rates for each k at various values of β and γ.

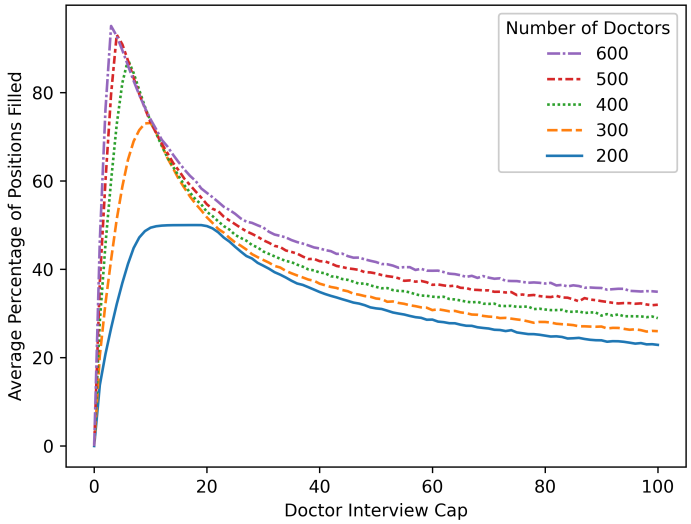


Figure 10: The average match rates for each k for different numbers of doctors when there are 400 hospitals.

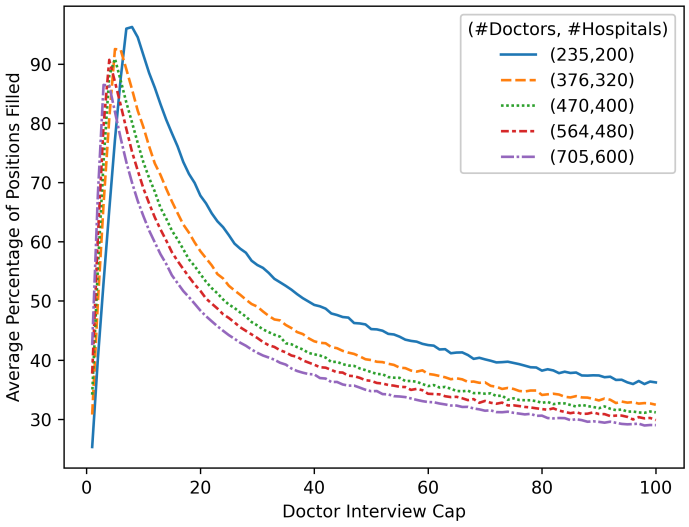


Figure 11: The average match rates for each k for different market sizes with the ratio of hospitals to doctors fixed at $\frac{400}{470}$.

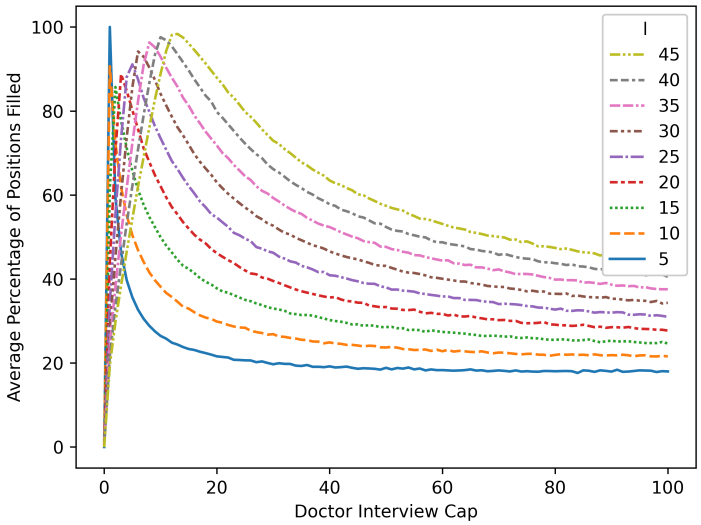


Figure 12: The average match rates for each k for different values of l, the hospitals' interview capacity.