Effect of Limited Number of Interviews on Matching Markets

HEDYEH BEYHAGHI, ÉVA TARDOS

We study the effect of the number of applications in the matching process modeling the national residency matching program (or applicants interviewing for jobs), where hospitals have a limit on the number of interviews and show non-intuitive effects in the matching outcomes:

- We show that when the number of interviews is limited, a market with limited number of applications results in a larger matching compared to a market with no limit.
- We show that a system of treating all applicants equally (setting the same limit for the number of applications for all) results in a larger matching than allowing a small set to apply to one more or one fewer positions. This comparison results in a scallop-shape function of expected size of matching with respect to expected number of applications.
- Finally we show that limiting the number of interviews can increase the size of the matching in certain cases.

1 INTRODUCTION

Matching is a fundamental paradigm in a variety of real-world situations and arises in various domains, e.g., students applying to attend schools, or colleges, applicants get matched with jobs at various job markets, and medical residents get matched with hospitals, just to name a few. In some of these domains, matchings are found through a centralized algorithm such as the deferred acceptance algorithm by Gale and Shapley [1962]. Prominent examples include: national residency matching program (NRMP) for assigning medical students to hospital residency programs; schools assignments in some cities the US; and college admissions in some countries [Roth and Sotomayor, 1992]. The deferred acceptance algorithm is typically studied assuming that participants express their full list of ordered preferences. Under this assumption, the algorithm finds a stable matching, meaning that no pair of participants would rather deviate from their assigned matches and be matched with one another. However, typically both sides of the match are limited in the number of options they can consider. In some systems, such as the national residency matching program, there are application limits imposed by the system, but even without such explicit limits, preparing applications, and interviewing applicants is costly and time consuming, and hence typically extremely limited. Applications, e.g. for college admission, often require writing specialized essays. Residency programs, as well as some of the colleges, interview their applicants, which requires significant time and effort. In large matching markets it is not feasible for either side to consider a large set. For instance, in NRMP, there are nearly 5000 residency programs and applicants have to limit their choices. Similarly, interviewing consumes significant time of the hospitals, and so hospitals can only grant interviews to a limited set of doctors. With a very limited number of interview and applications, one can view uncoordinated matching markets as roughly running a process analogous to deferred acceptance; applicants collect offers and as they turn down offers, new offers may be made.

The goal of this paper is to study the effect of the number of applications and the number of interviews on the resulting matching. We think of our model as a simplified version of the matching process of the national residency matching program or a job market process. In this program, the matching mechanism has two stages. In the first stage, doctors express
their interest in a set of hospitals, and hospitals choose to interview a subset of the applicants. In the second stage, both doctors and hospitals submit their ordered preference list over the set they interviewed with. The centralized matching market uses these preference lists to perform the deferred acceptance algorithm and outputs the final matching.

We will primarily consider a symmetric market, and as a primary metric of the social welfare of the system, we will mainly consider the size of the matching found. If there were no constraints on the number of interviews for applicants or programs, and both parties were willing to list everybody on the other side, the outcome of deferred acceptance algorithm would be a matching of maximum size (fully matching the smaller side of the market). With limited number of applications and interviews, the resulting matching can be much smaller. There are a number of different sources for this loss in the size of the matching. When doctors apply to a limited set of hospitals, there will be hospitals that have received less applications than their available positions, or even no applications at all. Even if hospitals receive many more applications than their positions, they might end up with free slots in case the doctors they interview choose other hospitals. Limiting the number of interviews by hospitals further reduces number of potential matches, e.g., doctors might end up with too few or no interviews, and even with interviews, they may not get an offer. An alternative view of the limitations of the hospital side is a limit on the offers for one position: based on interviewing all or none of the applicants, they may rank only a limited number of their favorite applicants. Such selective offer making is observed sometimes in academic job markets, when some high ranked departments prefer to wait till next year, rather than make offers to applicants they liked less.

The Model. We study the matching resulting from a two-stage procedure inspired by NRMP. In the first stage, applicants request interviews from a limited number of programs. We assume that applicants have a preliminary ranking of the hospitals (based on reputation), and will base their selection on this reputation. Next, the hospitals need to grant interviews to a select set of applicants and reject others. This again is based on a preliminary ranking of applicants based on their CV's. After conducting interviews, both applicants and hospitals can sort their interview set based on their refined preference list, and submit this list to the system running the deferred acceptance algorithm.

For the most part of this paper, we consider a model with the same number of doctors and hospital positions. In Section 3 we study a uniform and symmetric preference model where both doctors and hospitals have preferences drawn from uniform distributions, and after the interviews, they again have a preference list that is uniform. As we will show, with this assumption, each applicant applying to their top choices for hospitals, and both hospitals and applicants truthfully reporting their preferences, forms a Nash equilibrium of this game. In this model, reordering preferences after interviews does not change the distribution of preferences, so it is equivalent to think of the limitation of hospitals as a limit on the number of top applicants they are willing to rank, i.e., accept for a position.

In Section 5, we study the effect of limiting interviews in a tiered structure, where high tiers are preferred to low tiers but inside tiers the preferences are drawn from uniform distributions, as before. We will assume that the effect of the tiers is dominant, in that no party ever prefers a low tier option to a high tier option. In this model, both the applicants as well as the hospitals have an important strategic choice to make: how to divide the limited number of applications and interviews between the tiers. To study the effect of this limit on social welfare in a fair way, we have to model the values of low and high tier matches. We'll elaborate on this model further in Section 5.
Our Results. We study the effect of limitation on the number of applications and interviews in matching markets and show non-intuitive effects on the matching outcomes.

It seems intuitive that allowing applicants to apply to more positions increases the size of the resulting matching. In Section 3.1, we show that this is not true. We compute the expected size of the matching formally when each hospital selects only one applicant. In this special case, it is equivalent to view the process as hospitals immediately making an offer to the top applicant. When ranking of applications is correlated, such selective offer making is well-known to lead to many departments unable to fill their positions, as top applicants get many offers, and others don’t get any. Notice that we observe the same phenomenon in a model where rankings of different hospitals are independent, and drawn from the uniform distribution. Using simulation, we also show that the nonincreasing effect of additional applications remains when hospitals interview, and then rank more applicants, or make offers less selectively. By computing the expected size of matching, we find the optimal number of applications with which market achieves the largest matching. Our results suggest that when the number of interviews is low, it is best to also limit applicants to a low number of applications.

A different way to study the effect of additional applications is to start with a system where all applicants are allowed to apply to the same number of positions, and all hospitals interview the same number of candidates. Surprisingly, we find that this system of treating all applicants equally is best for maximizing the size of the matching! To show this, we allow a small subset of applicants to apply to one extra position. Surprisingly, we find that the increase in applications in this way decreases the size of the matching. On the other hand, if starting from the case when all doctors apply to the same number of hospitals, we restrict a small subset of doctors to one fewer application, this too decreases the size of the matching. In the case when all hospitals make an offer to their favorite applicant only, we show analytically that the equal treatment of all applicants is always preferable to favoritism or discrimination, by simply resulting in larger matching. For larger number of interviews, we establish the same phenomenon by simulation. The resulting effect on the size of the matching is illustrated by Figure 3 resulting in a scallop-shape function that represents the size of matching as a function of the expected number of applications.

In Section 4 we show the same phenomenon exists even in unbalanced markets, where the number of applicants and positions is different. In unbalanced markets, the optimum number of applications for social welfare (the size of the resulting matching) is different compared to the balanced case, and depends on the ratio between the two sides of the market. However, in this more general model, the effect of not treating everybody equally still can hurt the efficiency.

We also study the effect of different methods for distributing the applications, when the total number of allowed applications is fixed. We find the optimal method of distributing the applications when each hospital selects only one doctor. We show that in this case, the maximum matching is achieved when doctors either apply to $k$ or $k + 1$ positions; this allocation dominates any other way of distributing the applications, showing to maximize the size of the resulting matching, it is best to treat all applicants as evenly as possible.

In Section 5 we study the same phenomenon in a two-tier model: both doctors and hospitals are divided into high-tier and low-tier, and all parties prefer a match to a high-tier partner to a low-tier one. To evaluate social welfare in this model, we need to model this difference in value. We use a parameter $v$ to model the extent to which both sides value high-tier matches compared to low-tier. In this model, both applicants and hospitals need
to make a strategic choice about how to divide their applications and interviews between the two tiers. In the symmetric markets of Sections 3 and 4 increasing the number of interviews can only increase the size of the resulting matching. In the two-tier system modeling different qualities of applicants and positions in Section 5 we find that the social welfare can sometimes be improved with limiting the number of interviews.

**Related Literature.** While there has been great interest in finding stable matchings for various markets, there is less known about how limitation on the length of preference lists or number of interviews affect the market. Short preference lists were first studied by Immorlica and Mahdian [2005], however, they assumed that preference lists were inherently short, i.e. the participants preferred to stay unmatched to being matched outside their preference lists. In reality, the limited preference lists are mainly a product of the limitation explained above: limited time for information gathering and preparing applications, and interviewing applicants, etc. Similar to [Immorlica and Mahdian, 2005], limited applications on one side of the matching market were considered in various studies, including [Arnosti, 2016], [Kojima and Pathak, 2009].

Arnosti [2016] studies matching with short preference lists for applicants using a *large market approximation*, which we will also use in our paper. The large market approximation considers the limiting case where the number of applicants and positions grow while other parameters of the market, including the length of reported preference lists, remains fixed.

The limit on the number of applications due to limited budget or time, have been considered by Avery and Levin [2010] and Beyhaghi et al. [2017]. In these models, applicants need to be strategic about selecting a small set of positions to apply to. Particularly, when the market is correlated, applicants have a higher chance of admission in the less popular positions and they have to strategically come up with a short list balancing between maximizing their chance of admission in a high-ranked position and minimizing their chance of unemployment. Avery and Levin [2010] study early admission for college application. In early admission, that has become popular in recent years, students can select one college in early stage. They consider a model with two schools where students can choose between the more popular or the less popular school in the early stage. Beyhaghi et al. [2017], used a tiered model to study the case where the deferred acceptance algorithm is implemented with applicants submitting a limited list of applications and hospitals granting interviews to all their applicants. Their findings show how applicants decide where to apply and what the quality of outcome will be in the equilibrium. In this paper, we focus more on the social efficiency of outcomes, rather than outcome of strategic behavior. Compared to correlated preferences in these two papers, since we study independent preferences, the participants are not motivated to strategize their actions, and they only report their true preference lists. Therefore the loss in social welfare is not a result of strategic behavior. Also in our model both sides – and not only the applicants side – have limits on the length of preference lists.

There are also a few papers considering a two-stage game consisting of an interview stage, where both sides refine their preferences after the interviews. Drummond et al. [2016] and Kadam [2015] study the effect of limited number of applications on the matching in a two-stage game. In their models, both doctors and hospitals are (almost perfectly) vertically differentiated, so applicants and hospitals have almost perfect information on who they would match to, given full access to all parties preferences. While we agree that there are better and worse applicants, and better and worse residency programs, we believe that a lot of the choices are much more personal and idiosyncratic. To model this, we assume that
applicants’ preferences as well as hospitals’ preferences have a significant component that is random.

Drummond et al. [2016] investigate an interviewing game, where medical students and hospitals interview each other as the first stage of matching. For their theoretical analysis, they make the following assumptions: hospitals offering the residency programs have a common preference list; preference of medical students is distributed as a slightly different version of a common ordering; residents choose two hospitals to interview with; and hospitals grant interviews to all of their applicants. They show that under these assumptions, applicants select two consecutive hospitals in the common ordering that closely correspond to their own rank. Using simulation, they also show that with high dispersion, this is not equilibrium anymore and doctors select hospitals with less competition in this case. Similar to our model, they consider a two-stage game that involves an interview stage. Compared to their results, in the independent model that we study participants are not strategic and they only report their favorite choices. In light of this, we study the effect of limitation in number of applications without considering the effect of strategicness.

Kadam [2015] studies a two-stage model in which both doctors and hospitals have a limit on number of interviews. Similar to our paper, he shows that limiting the length of lists can have positive effects of the size of matching. However, these improvements have different reasons. In [Kadam, 2015], due to almost common preferences, some doctors are more preferred. With a stricter limit on the interviews for doctors, the more preferred doctors do not accept interviews with less preferred hospitals. Therefore in equilibrium hospitals grant interviews to less preferred doctors, which increases the probability for these doctors to be matched and increases the size of matching. However in our paper, in Sections 3 and 4 the preferences are independent and the improvement in the matching size is not a reason of popularity in the market. Also in Section 5, where the market is not symmetric, we show examples where both the size of matching and the social welfare increases with a limit on the number of interviews. However, in [Kadam, 2015] the increase in the size of matching occurs with a decrease in the social welfare.

Kleinberg et al. [2016] focus on interview as an information gathering step. They design an order for performing interviews and making offers that minimizes the number of interviews needed to find the stable matching. Although in their model, the positions are vertically differentiated, they assume that applicants do not know the perfect order prior to the process and the interviews help them to figure it out. Unlike our model where there is a limit on the number interviews due to the constraint, they continue the interviewing process until they find an efficient matching and study the overall number of interviews needed.

2 MODEL

We consider a two-sided market consisting of doctors and open positions in hospitals (hospitals for short). We assume there are $n$ doctors and $rn$ positions, and that each hospital has only one open position, so we focus on one-to-one matching.

We start with a model where there is a universal limit on the number of applications by doctors and the number of interviews by hospitals. The fixed limit on the number of applications by a doctor is inspired by NRMP that grants all doctors 10 initial applications. We relax this assumption later and compare the effect of allocating different number of applications to different doctors. We also have a fixed limit on the number of interview conducted by hospitals, modeling that hospitals spend more or less the same time and budget on interviews.
We consider a two-stage matching mechanism. In the first stage, doctors apply to a number of hospitals and ask for interview. Hospitals then choose a subset from their applications to conduct interviews. In the interview process, doctors and hospitals gather information and learn their preferences. In the second stage, at the end of interviews, both doctors and hospitals order the list that they interviewed with, based on their preferences. They submit their ordered preferences to the system which performs a doctor-proposing deferred acceptance algorithm to determine the final assignment of doctors to hospitals. The algorithm starts with all doctors unmatched. In each step, the algorithm simulates all unmatched doctors, who have not yet exhausted their options, propose to their most preferred hospital among those to which he/she has not yet proposed. Now the algorithm simulates that each hospital tentatively accepts their most preferred doctor from the doctors now proposing and the one who has been tentatively assigned to this hospital, and rejects all other doctors. The procedure is repeated until all unmatched doctors have been rejected from every hospital in their lists.

We use the following terminology throughout the whole paper.

**Application:** The procedure in the first stage where the doctors submit an application and ask for interview.

**Grant interview or reject for interview:** The procedure in the first stage where the hospitals grant interviews to a number of doctors that have applied for the position and reject the rest.

**Proposal:** The procedure in the second stage that simulates doctors proposing to hospitals as steps of the deferred acceptance algorithm.

**Offer:** In the special case where the hospitals limit their number of interviews to one, conducting interviews is of no use because the hospital offers the position to the selected doctor in any case. Therefore we use offer instead of interview in this case.

In the beginning of the procedure, everybody has prior preference over the other side. For Sections 3 and 4 we assume that everybody’s preference comes independently from uniform distribution over the other side. Both doctors and hospitals refine their preference order for the list they interviewed. However, we assume that the overall distribution of doctors preferences and hospitals preferences can be treated as outcome of new independent draws from uniform distribution. Refinement in the way mentioned, does not have an effect on our analysis. For analytic purposes we can assume that both the initial distribution, as well as the ranking after the interviews, are uniformly random.

We assume that throughout the whole process, there is no coordination between doctors or hospitals. If hospitals knew which applicants had more/less interviews, they may prefer to grant interviews to doctors with less interviews, as these are more likely to be matched with. Our model assumes no such strategic behavior.

For most of the paper we focus on the size of the matching as our notion of efficiency. We define social welfare as the ratio of size of the matching outcome to the size of the maximum matching. Since we assume all doctors/hospitals prefer to be matched to any hospital/doctor rather than being unmatched, in a maximum-size matching everybody on the less populated side of the market is matched.

**Definition 2.1 (Social Welfare).** We define social welfare of the matching outcome as the ratio of the size of the matching outcome compared to the size of the maximum-size matching. In a two sided market, maximum-size matching will be the size of the smaller side of the market.
Throughout this paper we use the "large market approximation" of the setting as introduced by [Arnosti, 2016]. This approximation assumes that the properties of the market, such as the limit on the number of applications, the limit on the number of interviews and the ratio between the size of the two sides of the market, is fixed while the size of the market grows to $\infty$, and studies the properties of the matching in the limiting case.

We claim that in this model, neither doctors or hospitals will be strategic: When doctors apply to hospitals for interview, they pick their favorite hospitals, with respect to their prior preference. Also hospitals select their favorite doctors to interview based on their prior. After interviews doctors and hospitals refine their preference order based on what they learned in the interviews stage. They are truthful with the orders they use for doctors/hospitals they interviewed. To see why, note that in the first stage due to the symmetry and independence in the market, no doctor or hospital is considered more popular than others. Therefore neither doctors or hospitals can benefit from applying or granting interviews to somebody other than their favorite selection. After the interviews, again neither hospitals or doctors can benefit by submitting their lists in a non-truthful order. It is not beneficial for the doctors (the proposing side) as shown in [Gale and Shapley, 1962]. Also as shown in [Immorlica and Mahdian, 2005], the probability that it is beneficial for hospitals vanishes as the size of the market grows to $\infty$, which is the setting we study in this paper. We summarize the statement in the following lemma.

**Lemma 2.2.** Both doctors and hospitals are truthful during the matching procedure:

*In the first stage, doctors apply to their favorite hospitals; and hospitals grant interviews to their favorite applicants.*

*In the second stage, both doctors and hospitals submit the list that they interviewed, ordered from the most preferred to the least preferred.*

### 2.1 No Limit for Granting Interviews

Before studying the effect of the limit on the number of interviews by hospitals, we recall the analysis for the baseline case, studied in our previous paper [Beyhaghi et al., 2017] with $n$ doctors and $n$ hospitals in the market. In this case there is no limit on the number of interviews by hospitals. We considered the game when doctors are allowed to list only up to $k$ hospitals, for $k << n$.

**Example 2.3.** We show when $k = 1$, the social welfare of the matching is $(1 - 1/e) \approx 0.63$ in equilibrium. In a matching between hospitals and doctors the size of matching equals the number of matched hospitals, therefore the social welfare the ratio of the number of matched hospitals to the size of maximum matching which equals the fraction of matched hospitals. As Lemma 2.2 implies, each doctor applies to exactly one hospital who receive an application in the first stage receives a proposal and become matched.

When doctors apply to more than one hospital ($k > 1$), Arnosti [Arnosti, 2016] proposes the following formula for the probability $p$ of a single proposal resulting to a permanent match as $n$ grows to $\infty$. The main idea is that in the limit as $n$ goes to infinity, the probability of a proposal resulting in a match is independent of the previous proposals of the applicant being rejected.
Proposition 2.4. [Arnosti, 2016] Using the above large market approximation with each doctor applying to $k$ hospitals, the probability $p$ of a single proposal resulting in a permanent match in the deferred acceptance procedure, satisfies the following equation.

$$(1 - p)^k = e^{-(1-(1-p)^k)/p} \quad (1)$$

Proof idea. Since the outcome of the deferred acceptance algorithm does not depend on the order in which doctors propose, we may hold out a single doctor $d$ and run the deferred acceptance algorithm on the remainder of the market. Now consider $d$ proposing to her favorite position. Her first few proposals may get rejected. Once a hospital accepts her proposal, it may reject a different doctor, who may propose for her next position, etc. We call the resulting sequence of rejections a rejection chain. The probability that a proposal of doctor $d$ causes a rejection chain that gets doctor $d$ rejected from the hospital that first accepted her vanishes as the market grows, therefore we may assume that $d$’s first tentatively accepted proposal will lead to a permanent match. Also, in a large market, learning that $d$’s first $m$ proposals got rejected does not affect the probability of acceptance of other proposals. Thus, from $d$’s perspective, each hospital that she applied to in the first stage, should be available to her with some probability $p$, and their availability should be independent. With this argument, the probability that $d$ matches is $1 - (1 - p)^k$, and the expected number of hospitals $d$ proposes to is

$$1 + (1 - p) + (1 - p)^2 + ... + (1 - p)^{k-1} = \frac{1}{p}(1 - (1 - p)^k)$$

From the point of view of each hospital, each of these proposals is sent to them roughly with probability $1/n$; thus the probability that a hospital receives at least one proposal is

$$1 - (1 - 1/n)^k(1-(1-p)^k) \approx 1 - e^{-(1-(1-p)^k)/p}.$$ 

Since doctors match with probability $1 - (1 - p)^k$, and the number of doctors and hospitals that match must be equal, we have that

$$(1 - p)^k = e^{-(1-(1-p)^k)/p}.$$ 

Figure 1 shows the expected size of matching as a function of the expected number of applications by doctors. We extended the figure for a fractional number of applications. When the expected number of applications is $x$, such that $k < x < k + 1$, doctors have either $k$ or $k + 1$ applications. As seen in the figure, the matching size is increasing in the number of applications.

Lemma 2.5. Suppose that either all the doctors apply to the same number of hospitals or a fraction of them apply to $k$ while others apply to $k + 1$ hospitals. In a setting where there is no limit on the number of interviews by hospitals and the preferences are uniformly random, the social welfare of the matching is increasing in the expected number of applications.

Proof. We show that if one of the doctors who previously had $k$ applications, now has $k + 1$ applications and the number of applications of other doctors remain the same, the size of the matching can only increase. By lemma 2.2, both doctors and hospitals are truthful, therefore we are comparing the size of matching as the result of deferred acceptance algorithm when doctor $d$ has an extra application and all other doctors have the same number of applications. Since the deferred acceptance algorithm is oblivious to the order in which doctors propose, we may hold out the last application of doctor $d$ and find the
outcome when the doctor $d$ does not have this application in his/her list. The result of the deferred acceptance algorithm, without this application is the same as the case where doctor $d$ had $k$ applications. We show that this last application can only increase the number of matching. If doctor $d$ is matched with one of his/her first $k$ proposals, the last application does not change. If doctor $d$ proposes to $k+1_{st}$ hospital, it can only increase the number of hospitals who have received any proposal.

In Section 3, we evaluate the size of matching as a function of number of applications with different number of offers or interviews by hospitals. We assume that the market is symmetric in the sense that doctors apply to the same (closely bounded) number of applications. In Section 4, we study unbalanced networks where the set of doctors and hospitals are not the same size. In Section 5, we study the effect of interviews in a more general model, explained later; and we show that limiting the number of interviews, does not always hurt the social welfare, but it can improve it in some cases.

### 3 EFFECT OF NUMBER OF INTERVIEWS

In this section we study the effect of number of interviews/offers with the same number of doctors and hospitals in the market. In 3.1 and 3.2 we find the efficiency of matching as a function of the expected number of applications by doctors, when hospitals only select one applicant in the first stage of the mechanism. We begin with the case where all doctors apply to the same number of hospitals and then we move to the case where doctors send different number of applications. We find allocating the number of applications equally results in the most efficient matching and allowing a small subset to apply to an extra position or restricting a small set to apply to one less position, results in a smaller matching. In Section 3.3 we study the matching outcome when hospitals grant multiple interviews/offers. We find that treating applicants equally is most efficient even in this more general model.

#### 3.1 Same Number of Applications and Making One Offer

In this part we assume that hospitals only select one of their applicants in the first stage of the matching. As shown in Lemma 2.2, they choose their favorite applicant. Since interviews are conducted to compare the selected doctors, in this case that only one doctor is selected,
interviews are of no use from the hospital’s point of view. Therefore the hospital immediately offers the position to the selected doctor. On the other hand, the outcome of the deferred acceptance algorithm can be easily found in this case; each doctor will be matched to his/her favorite hospital among those who gave an offer. Therefore without going through the complication of the deferred acceptance algorithm, we assume that each doctor accepts the most preferred offer. This discussion implies that the size of matching in this case, is the number of doctors who receive an offer.

We study the social welfare of the matching with respect to the number of applications by doctors. First, consider the simple case where doctors are allowed to send one application. Since in this case each doctor receives at most one offer, all the offers are accepted and the size of matching is equal to the number of offers which is the same as the number of hospitals who received an application. As shown in example 2.3, because of the uniformly random preferences, the size of matching in this case tends to $1 - \frac{1}{e}$ in the limit as $n$ grows to $\infty$.

Now to observe the effect of having more applications, consider the extreme case where doctors have no limit and are allowed to apply to all hospitals. If the size of matching is monotone increasing in the number of applications, this case creates the largest matching. Since hospital preferences are uniformly random, each hospital selects an applicant to make an offer to, uniformly at random. Therefore the number of doctors who receive an offer is:

$$\lim_{n \to \infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - \frac{1}{e}$$

which shows that although the number of applications increased from one to $n$, there is no increase in the size of matching. Therefore either $1 - \frac{1}{e}$ is the largest matching possible or applying to $1 < k < n$ positions, achieves a larger matching which, if true means that matching size is not increasing in the number of applications.

**Definition 3.1 (covered hospital).** A hospital is covered if it receives at least one application.

Suppose that all doctors are allowed to apply to $k$ positions, therefore the number of total applications is $nk$. We show the number of covered hospitals is $\approx n(1 - \frac{1}{e} - \frac{1}{k})$.

**Proposition 3.2.** When all doctors apply to exactly $k$ uniformly random hospitals, with large market approximation the expected fraction of covered hospitals is $(1 - \frac{1}{e} - \frac{1}{k})$.

**Proof Sketch.** Similar to example 2.3, the probability of a hospital receiving any application is $\approx 1 - (1 - 1/n)^{nk}$, which tends to $(1 - e^{-k})$ in the limit. \(\square\)

Next we find the probability of a random application turning into an offer. In Lemma 2.2 we showed that doctors apply to their favorite positions. Since doctor preferences are independent, applications of different doctors are also independent. Also since with large $n$ and small number of applications, the probability that a fresh truly random choice of a doctor is identical to a previous choice approaches to 0, with large market assumptions, we assume that different applications of a doctor are independent. Therefore we can assume that the destination hospitals of all applications are selected uniformly and independently at random. On the other hand doctors have no knowledge about the number of doctors applying to each hospital and all hospitals look symmetric to them. Thus from every doctors perspective, each hospital that they applied to, makes an offer to them with some probability $p$, and this probability is independent among the applications.

**Proposition 3.3.** When each doctor applies to $k$ random positions and each hospital offers its position randomly to one of its applicants, the expected social welfare of matching
equals

\[ 1 - \left(1 - \frac{(1 - e^{-k})}{k}\right)^k. \]

**Proof.** With the above arguments we know from the doctors perspective, each hospital that they applied to, offers them a position independently with probability \( p \). Therefore the expected number of total offers is the product of \( p \) and the total number of applications. As discussed previously the total number of offers equals the number of covered hospitals and is equal to \( n(1 - e^{-k}) \). Also since each doctor applies to \( k \) positions, the total number of applications is \( nk \). Therefore \( p = \frac{1 - e^{-k}}{k} \) and the probability for a doctor to receive an offer is \( (1 - (1 - \frac{(1 - e^{-k})}{k}))^k \). The expected social welfare of matching is equal to the probability of a doctor receiving an offer, which implies the conclusion. \( \square \)

Figure 2 shows the size of matching with respect to the number of applications.

![Fig. 2. Size of matching with respect to the number of applications, when hospitals make only one offer.](image)

As seen in Figure 2 the maximum size of matching occurs when all doctors apply to 3 hospitals. Therefore from a market design point of view, allowing three applications in this case is optimal in terms of social welfare.

**Proposition 3.4.** In markets with the same number of applicants as positions where participants’ preferences are drawn independently and uniformly at random, if all doctors are allowed to apply to \( k \) positions and hospitals grant a single offer, allowing doctors to apply to 3 hospitals results in the largest matching outcome when size of the market grows to \( \infty \).

As shown in Figure 2, having more applications can have both positive and negative effects: As a positive effect, with more applications the number of hospitals who receive an application increases. This can potentially increase the size of matching. On the other hand as the number of applications increases, the hospitals become more congested. Since the total number of offers is limited by the number of hospitals (each hospital makes at most one offer), this causes an increase in the probability of rejection of an application. This may increase the number of doctors with no offer, leading to a negative effect on the size of matching. As seen in the picture, The biggest jump in the size of matching occurs when moving from one application to two applications. When increasing the number of applications from one to two, the fraction of covered hospitals changes from \( 1 - 1/e \approx 0.63 \) to \( 1 - 1/e^2 \approx 0.86 \), which is the highest increase when adding more applications. This causes
the highest increase in the matching size. The increase stops with only three number of applications, with which the fraction of covered hospitals is $1 - 1/e^3 \approx 0.95$. From this point forward the negative effect takes over and because of the random allocation of hospitals some doctors receive multiple offers while others receive none.

### 3.2 Fractional Expected Number of Applications and Making One Offer: scallop-shape

Unlike Section 3.1, where there was a universal limit on the number of applications by doctors, in this section we study the social welfare when doctors are allowed to send different number of applications. First we show that for any expected number of applications $x$, the optimal social welfare occurs when doctors send either $\lceil x \rceil$ or $\lfloor x \rfloor$ applications. Then we study the social welfare as a function of the expected number of applications. Surprisingly, we observe that granting extra applications to a small set of doctors and also retracting an application from a small set both hurt the market; suggesting that unfair treatment is not efficient in terms of social welfare.

The following theorem shows that for any expected number of applications $x$, the optimal social welfare occurs when doctors send either $\lceil x \rceil$ or $\lfloor x \rfloor$ applications.

**Lemma 3.5.** With any expected number of applications, $x$, the distribution of number of applications that achieves the highest social welfare, is one that allocates $\lfloor x \rfloor$ applications to some doctors and $\lceil x \rceil$ to the others, such that the expected number of applications equals $x$.

**Proof.** Suppose the applications are distributed in a different way. Therefore there are two doctors with $l$ and $k$ applications such that $k - l \geq 2$. We show that the size of matching is improved if we allocate $f = \lfloor (k + l)/2 \rfloor$ and $c = \lceil (k + l)/2 \rceil$ applications to those doctors. This alteration does not change the probability of receiving offers by other doctors as the applications are independent from doctors perspective. So the only difference is the probability of receiving offers by these two doctors. Let $p$ be the probability of a random application to lead to an offer. We claim

$$1 - (1 - p)^k + 1 - (1 - p)^l \leq 1 - (1 - p)^c + 1 - (1 - p)^f.$$  

Since $(1 - p)^c$ and $(1 - p)^f$ have the same product as $(1 - p)^k$ and $(1 - p)^l$, their sum is larger when the two factors are far apart, therefore:

$$(1 - p)^c + (1 - p)^f \leq (1 - p)^k + (1 - p)^l$$

which implies the conclusion. 

This argument shows that in order to find the optimal social welfare for different expected number of applications, we only need to study the case where doctors apply to the same number of hospitals—as studied in 3.1— or two consecutive numbers. To find the size of matching, we first find the number of hospitals who receive an application.

**Proposition 3.6.** The fraction of covered hospitals, when the expected number of applications is $x$, equals to $(1 - e^{-x})$.

**Proof.** Similar to Proposition 3.2.

Using the proposition above, we are ready to find the social welfare with different expected number of applications.
Theorem 3.7. The social welfare of the matching with expected number of applications $k < x < k + 1$, when applicants either apply to $k$ or $k + 1$ positions is:

$$([x] - x)(1 - (1 - \frac{1-e^{-x}}{x})^k) + (x - [x])(1 - (1 - \frac{1-e^{-x}}{x})^{k+1})$$

This function is illustrated in Figure 3.

Proof. As shown in the proof of Proposition 3.3, from the doctors perspective, the probability of a random application, leading to an offer is:

$$\frac{\text{number of covered hospitals}}{\text{number of applications}} = \frac{1-e^{-x}}{x}.$$ 

Therefore the expected social welfare of the matching which is the same as the probability of a random doctor receiving an offer is:

$$([x] - x)(1 - (1 - \frac{1-e^{-x}}{x})^k) + (x - [x])(1 - (1 - \frac{1-e^{-x}}{x})^{k+1}).$$

Fig. 3. Size of matching with respect to expected number of applications, when hospitals make only one offer.

The scallop-shape figure 3 illustrates the size of matching as a function of expected number of applications. The unusual behavior of the function at integer points shows that allowing a small group to apply to one more position, or limiting the number of applications of a small group to one less application, has a negative effect on size of matching. Also, if we compare this diagram with figure 1 as both use the same method for distributing applications; i.e. same number of doctors applying to $k$ or $k + 1$ positions, we observe that this unusual shape belongs to the case with limited number of interviews, and if there is no limit on the number of interviews the size of the matching is monotone increasing.

As seen in Figure 4 the maximum size of matching occurs when all doctors apply to 3 hospitals, when doctors either apply to the same number of hospitals or two consecutive numbers. Also as shown in Lemma 3.5, for any expected number of application this is the optimal distribution of applications in terms of social welfare. This implies the following proposition which is a generalization of proposition 3.4: Applying to 3 hospitals is not only optimal when all doctors apply to the same number of positions, but it is also optimal for any arbitrary allocation of applications to doctors.

Proposition 3.8. In markets with the same number of applicants as positions where participants’ preferences are drawn independently and uniformly at random, allowing doctors
to apply to 3 hospitals results in the largest matching outcome when size of the market grows to $\infty$.

3.3 Granting multiple interviews

In this section we focus on the more realistic case, where hospitals grant multiple interviews and we observe that the size of matching as a function of expected number of applications has a similar structure. Unlike the previous subsections where each hospital immediately made an offer to their top applicant, in this case they grant multiple interviews and have backup options in case they do not match their top applicant.

With multiple interviews, finding the final matching requires running the usual procedure of the deferred acceptance algorithm and finding the size of matching is more complicated. In the previous case where hospitals made immediate offers to their top applicant, each doctor was matched to their most preferred hospital that they received an offer from. However, when hospitals interview multiple doctors this is not clear anymore. A doctor can be rejected by a hospital in the first stage before the interview, if the hospital does not grant an interview to the doctor. Also a doctor can be rejected in the second stage as a step of the deferred acceptance algorithm. With these different potential outcomes of an application, the probability of an application leading to an offer is more complicated than the cases previously studied.

Due to the theoretically more complicated scenario, we used simulation to find the resulting matching size when doctors grant multiple interviews. As an example, we considered a market with 10,000 hospitals and doctors and simulated the outcome of the matching procedure. The red curve in Figure 5, shows the size of matching with respect to expected number of applications when hospitals grant two interviews.

As the figure shows, the size of matching even with multiple interviews has the same structure as a single interview (the blue curve) and more interviews results in a larger matching. In the case with single interview hospitals reject more applications in the first stage. This elimination results in shortened lists as inputs to the deferred acceptance algorithm. Therefore it is not surprising that more interviews results in larger matchings.

4 UNBALANCED MARKETS

In this section we show that the phenomenon of the positive effect of setting the same limit for all doctors is not limited to a balanced market – where the number of applicants and positions are the same – but extends to unbalanced markets with different number of
applicants and positions. Although the same phenomenon exists more generally, the exact function is not preserved and the optimum number of applications depends on the ratio between the sizes of the two sides.

We study the case where hospitals make one offer and the ratio of the number of hospitals to the number of doctors is $r$. Similar to the previous section, we can compute the size of the matching as a function of the expected number of applications, when doctors send $k$ or $k + 1$ applications.

**Proposition 4.1.** The expected fraction of matched doctors with expected number of applications $k < x < k + 1$, when applicants either apply to $k$ or $k + 1$ positions and when the ratio of number of hospitals to number of doctors is $r$ is:

$$(\lceil x \rceil - x)(1 - \frac{r(1 - e^{-x/r})}{x})^k + (x - \lfloor x \rfloor)(1 - \frac{r(1 - e^{-x/r})}{x})^{k+1}).$$

**Proof.** As shown in the proof of Proposition 3.3, from the doctors perspective, the probability of a random application, leading to an offer is $\frac{\text{number of covered hospitals}}{\text{number of applications}}$. From the hospitals perspective, each application is equally likely to be sent to each hospital. Therefore the number of applications received is distributed as a Poisson distribution with $\lambda = \frac{x}{r}$. So the number of expected covered hospitals is $rn(1 - e^{-x/r})$ and the probability of a random application, leading to an offer is $\frac{r(1 - e^{-x/r})}{x}$. Therefore the expected fraction of doctors who are matched which is the same as the probability of a random doctor receiving an offer is:

$$(\lceil x \rceil - x)(1 - \frac{r(1 - e^{-x/r})}{x})^k + (x - \lfloor x \rfloor)(1 - \frac{r(1 - e^{-x/r})}{x})^{k+1}).$$

Based on Definition 2.1, the social welfare of a matching is the ratio of the size of the matching to the size of maximum matching and in unbalanced markets, the maximum size is the size of the smaller side of the matching. For $r \geq 1$, the doctors make the smaller side therefore the social welfare is equal to expected fraction of doctors who are matched as computed in Proposition 4.1. If $r < 1$, the social welfare is the expected fraction of doctors who are matched as computed in Proposition 4.1 divided by $r$. 

![Fig. 5. Size of matching with respect to expected number of application. The top red curve represents the setting where hospitals allow two interviews. The lower blue curve represents the setting with one interview and is re-drawn from Figure 3 for reference.](image-url)
Figure ?? shows the social welfare of the matching as a function of expected number of applications for $r = \frac{1}{2}, 1, 2$. In the figure, the red function refers to $r = 2$, the blue function to $r = 1/2$ and the green function to $r = 1$.

![Fig. 6. Social welfare of the matching with respect to the expected number of applications. The red curve refers to $r = 2$, blue curve to $r = 1/2$ and the green curve to $r = 1$, where $r$ is the ratio of the number of hospitals to doctors.](image)

As seen in the figure a similar structure holds for unbalanced networks, but the optimal number of applications depends on the factor of balancedness $r$. When the number of hospitals is half of the number of doctors, the market achieves its maximum size when each doctor applies to just one position. With more applications, hospitals become more congested, and the number of covered hospitals and therefore the number of total offers does not increase significantly. Therefore the rejection probability of applications increases and with higher probability a doctor remains unmatched. In contrast, when the number of hospitals is more than the number of doctors, allowing more applications has a positive effect. With more applications –while the number of applications is still small– more hospitals receive at least one application; therefore the number of offers increases considerably.

Also note that in the figure, the social welfare of the balanced market is generally lower than the social welfare of markets with $r = 1/2, 2$. This is not surprising since by definition the social welfare is the fraction of matched individuals on the smaller side of the market: doctors for $r \leq 1$, and hospitals for $r > 1$. For $r = 2$, with the same number of applications, more hospitals are covered that leads to more offers and a higher fraction of matched doctors. For $r = 1/2$, with the same number of applications a higher fraction of hospitals is covered which leads to a higher fraction of matched hospitals.
5 IMPROVEMENT OF SOCIAL WELFARE WITH RESTRICTING INTERVIEWS

In this section we study the effect of limiting the number of interviews in a more general model. Beyhaghi et al. [2017] study a setting where the preference model is a combination of independence and correlation. They assume that both sides of the market are divided into tiers, where everybody prefers higher tiers to lower tiers (correlation) and inside a tier the preferences are uniformly random (independence). It is not hard to see that the model studied in the previous sections is a special case of the tiered model where the market only consists of one tier. In the model studied in the previous sections, limiting the number of interviews by hospitals had a negative effect on the social welfare as observed in Figure 5 which compares one and two interviews. This holds more generally since with a more restricted limit on the number of interviews by hospitals, more applications are rejected in the first stage of the process. Therefore doctors and hospitals submit shortened lists to be used by the deferred acceptance procedure, resulting in a matching with smaller size. Our main finding in this section is that a limit on interviews can improve the social welfare in the multi-tiered model.

In a multi-tiered market, matching to a higher tier doctor/hospital has a higher value for a participant compared to a lower tier, and the social welfare is the sum of the values of all participants (doctors and hospitals). For example in a two-tier market, matching to a high-tier doctor/hospital has value \( v > 1 \) for participants, while matching to a low-tier doctor/hospital has value 1.

**Definition 5.1 (social welfare in a two-tier market [Beyhaghi et al., 2017]).** Social welfare is the sum of the value of matched doctors and hospitals, such that the value of matching to a high-tier doctor/hospital is \( v \) while the value of matching to a low-tier doctors/hospitals is 1.

In a tiered market, the participants are strategic and their strategic action can hurt the social welfare [Beyhaghi et al., 2017]. Consider a two-tier market. Unlike the model in the previous section, since the market is not symmetric and high-tier hospitals are more popular, doctors are strategic. They need to decide how to select their limited number of hospitals both to increase the likelihood of being matched to a high-tier hospital, and to have backup options from the low tier in case they are not matched to the high tier. In the equilibrium of a two-tier market, doctors have a tendency to apply more to high-tier hospitals compared to a non-strategic setting when their goal is to maximize the social welfare [Beyhaghi et al., 2017]. This strategic behavior sometimes has a negative impact on social welfare of the matching. Applying to more high-tier hospitals makes high-tier hospitals congested and also reduces the number of matches of the high-tier doctors (more valuable doctors in terms of social welfare) since they do not use the benefit from potential matches to the low tier.

Although restricting the number of interviews hurts the social welfare in many cases, there are settings like Example 5.2 where it can improve the social welfare.

**Example 5.2.** Consider a market with the same number of top-tier doctors, low-tier doctors, top-tier hospitals and low-tier hospitals. Suppose each doctor sends two applications and \( v, \) the value for the high tier, is \( \approx 1.91. \) The social welfare of the matching when each hospital grants at most six interviews is higher compared to case where there is no limit on the number of interviews.

The result in Example 5.2 is found via simulation. We observed \( v \approx 1.91 \) is the smallest value that when there is no limit on the number of interviews, in equilibrium all high-tier
doctors send both of their applications to the high tier. When hospitals grant a limited number of interviews, the probability of an application leading an offer decreases. So high tier doctors gain less utility when sending both of their applications to the high tier and have more incentives to apply to the low tier, therefore they randomize between sending both applications to the high tier or one to the high and one to the low tier. By simulation we observed that when the number of interviews by each hospital is limited to six, this new flow of applications from high tier doctors to low tier hospitals increases the size of the matching and also improves the overall social welfare. With less number of interviews, many applications are rejected in the first stage and the social welfare decreases. Also with more number of interviews the flow of applications to the low tier is not significant to improve the social welfare. Therefore we observe that in this case a limit on the number of interviews by hospitals (six) results in a more efficient matching than unlimited number of interviews.

REFERENCES


