TWO-SIDED COMPETING MATCHING MARKETS WITH COMPLEMENTARY PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a new algorithm for addressing the problem of two-sided matching markets with complementary preferences, where agents' preferences are unknown a priori and must be learned from data. The presence of complementary preferences can lead to instability in the matching process, making this problem challenging to solve. To overcome this challenge, we formulate the problem as a bandit learning framework and propose the Multi-agent Multi-type Thompson Sampling (MMTS) algorithm. The algorithm combines the strengths of Thompson Sampling for exploration with a double matching technique to achieve a stable matching outcome. Our theoretical analysis demonstrates the effectiveness of MMTS as it is able to achieve stability at every matching step and has a sublinear Bayesian regret over time. Our approach provides a useful method for addressing complementary preferences in real-world scenarios.

1 INTRODUCTION

Two-sided matching markets have been a mainstay of theoretical research and real-world applications for several decades since the seminal work by (Gale and Shapley, 1962). Matching markets are used to allocate indivisible "goods" to multiple decision-making agents based on mutual compatibility as assessed via sets of preferences. Matching markets embody a notion of scarcity in which the resources on both sides of the market are limited. One of the key concepts that contribute to the success of matching markets is *stability*, which criterion ensures that all participants have no incentive to block a prescribed matching (Roth, 1982). Matching markets often consist of participants with *complementary* preferences that can lead to instability (Che et al., 2019). Examples of complementary preferences in matching markets include: firms seeking workers with skills that complement their existing workforce, sports teams forming teams with players that have complementary roles, and colleges admitting students with diverse backgrounds and demographics that complement each other. Studying the stability issue in the context of complementary preferences is crucial in ensuring the successful functioning of matching markets with complementarities.

In this paper, we propose a novel algorithm and present an in-depth analysis of the problem of complementary preferences in matching markets. Specifically, we focus on a many-to-one matching scenario and use the job market as the example. In our proposed model, there are a set of agents (e.g., firms), each with limited quota, and a set of arms (e.g., workers), each of which can be matched to at most one agent. Each arm belongs to a unique type, and each agent wants to match with a minimum quota of arms from each type. This leads to complementarities in agents' preferences. Additionally, the agents' preference of arms from each type is unknown a priori and must be learned from data, which we refer to as the problem of *competing matching under complementary preference* (CMCP).

Our first result is the formulation of CMCP into a bandit learning framework as described in (Lattimore and Szepesvári, 2020). Using this framework, we propose a new algorithm, the Multi-agent Multi-type Thompson Sampling (MMTS), to solve CMCP. Our algorithm builds on the strengths of Thompson Sampling (TS) in terms of exploration and further enhances it by incorporating a *double matching* technique to find a stable solution under CMCP. The TS algorithm, as described in (Thompson, 1933; Agrawal and Goyal, 2012; Russo et al., 2018), can effectively address the *incapable exploration* problem in the competing matching problem, as described in (Liu et al., 2020), by using the randomized sampling, also illustrated in Section 3.2. Unlike the upper confidence bound (UCB) algorithm, TS method can achieve sufficient exploration by incorporating a deterministic, non-negative bias inversely proportional to the number of matches into the observed empirical means. Furthermore, the double matching technique proposed in this paper uses two stages of matching to satisfy both the type quota and total quota requirements. These two stages mainly consist of using the deferred acceptance (DA) algorithm from (Gale and Shapley, 1962), which is easy to be implemented.

Second, we present a theoretical analysis of the proposed MMTS algorithm. Our analysis shows that MMTS can achieve stability at each matching step and show the incentive compatibility (IC) of the MMTS. The proof of stability is obtained through a two-stage design of the *double matching* technique, and the proof of IC is obtained through the lower bound of the regret. To the best of our knowledge, MMTS is the first algorithm to achieve stability and IC in the CMCP.

Finally, our theoretical results indicate that MMTS can achieve a Bayesian total regret that scales with the square root of the time horizon (T) and is nearly linear in the total quota of all firms (Q). Furthermore, we find that the Bayesian total regret only depends on the square root of the *maximum* number of workers ($K_{\max}$) in one type rather than the square root of the total number of workers ($\sum_m K_m$) in all types. This is a more challenging setting than that considered in previous works such as (Liu et al., 2020; Jagadeesan et al., 2021), which only consider a single type of worker in the market and a quota of one for each firm. To address these challenges, we use the eluder dimension (Russo and Van Roy, 2013) to measure the uncertainty set widths and bound the instantaneous regret for each firm, and use the union bound of concentration results to measure the probability of *bad events* occurring to get the final regret. Bounding the uncertainty set width is the key step for deriving the sublinear regret upper bound of MMTS.

The rest of this paper is organized as follows. In Section 2, we introduce the necessary components in the problem of CMCP. Meanwhile, we also state the challenges of this problem. In Section 3, we provide the MMTS algorithm, its comparison with other algorithms, and show the incapability of the UCB algorithm in CMCP. Then we present the stability, regret upper bound, and the incentive-compatibility of the MMTS in Section 4. Finally, in Section 5, we show two examples, present the distribution of learning parameters, and demonstrate the robustness of MMTS in large markets.

2 Problem

2.1 Problem Formulation

We now describe the problem formulation of the **C**ompeting **M**atching under **C**omplementary **P**references problem (CMCP). Using the scenario of worker-firm matching as our running example, we introduce the notation and key components of the CMCP. We define T as the time horizon and without loss of generality, we assume it is known[1]. We denote $[N] = [1, 2, ..., N]$ where $N \in \mathbb{N}^+$. Define the bold $\mathbf{x} \in \mathbb{R}^d$ be a d-dimensional random vector.

(I) Environment. We consider a centralized platform with N firms, denoted by the set $\mathcal{N} = \{p_1, p_2, ..., p_N\}$, and various types of workers, represented by sets $\mathcal{K}_m = \{a_1^m, a_2^m, ...a_{K_m}^m\}$, for $m \in [M]$, where K_m is the number of m-th type workers. Each firm p_i has a specific minimum type quota q_i^m to recruit m-type workers, and a maximum total quota Q_i (e.g., seasonal headcount in company), for all $i \in [N], m \in [M]$ and we assume $\sum_{i=1}^M q_i^m \leq Q_i$. Additionally, we define the total market quota as $Q = \sum_{i=1}^N Q_i$ and the total number of available market workers as $K = \sum_{m=1}^M K_m$. It is assumed without loss of generality that the total number of quotas is greater less than the total number of available market workers ($Q \ll K$) and T is large.

(II) Preference. We give preferences of both sides of the market. There are two preference lists: the preferences of workers towards firms, and the preferences of firms towards workers.

a. Preferences of m-type workers towards firms $\boldsymbol{\pi}^m : \mathcal{K}_m \mapsto \mathcal{N}, \forall m \in [M]$. We assume that preferences of types of worker to firms are fixed, known over time. For instance, workers submit their preferences over different firms to the platform. We denote $\pi_{j,i}^m$ as the rank order of firm p_i in the preference list of $m-$type worker a_j^m, and assume that there are no ties in the rank orders[2]. The centralized platform knows the fixed preferences of $m-$type worker towards firms, denoted as

[1] The unknown T can be handled with the well-known doubling trick (Auer et al., 1995).

[2] The strict preference is not necessary for stable matching and regret metric. However, for simplicity, we assume preference is strict, which avoids the multiple optimal stable matching solutions even they are equivalent.

$\boldsymbol{\pi}_j^m \subseteq \{\pi_{j,1}^m, ..., \pi_{j,N}^m\} \cup \{a_j\}, \forall a_j \in \mathcal{K}_m$ and $m \in [M]$, and singleton a_j represents the worker's preference to remain unmatched. In other words, $\boldsymbol{\pi}_j^m$ is a subset of the permutation of $[N]$ plus the worker itself. And $\pi_{j,i}^m < \pi_{j,i'}^m$ implies that $m-$ type worker a_j^m prefers firm p_i over firm $p_{i'}$ and as a shorthand, denoted as $p_i <_j^m p_{i'}$. This known worker-to-firm preference is a mild and common assumption in the matching market literature (Liu et al., 2020; 2021; Li et al., 2022).

b. Preferences of firms towards m − type workers $\mathbf{r}^m : \mathcal{N} \mapsto \mathcal{K}_m, \forall m \in [M]$. The true *unknown* preferences of firms towards workers are fixed over time. The goal of the platform is to infer these unknown preferences through historical matching data. We denote $r_{i,j}^m$ as the true rank of worker a_j^m in the preference list of firm p_i, and assume there are no ties. p_i's preferences towards workers is represented by $\mathbf{r}_i^m$, which is a subset of the permutation of $[\mathcal{K}_m]$ plus the firm p_i itself, representing the firm's preference to remain unmatched. Here $r_{i,j}^m < r_{i,j'}^m$ implies that firm p_i prefers worker a_j^m over worker $a_{j'}^m$, and the notation $a_j^m <_i^m a_{j'}^m$ similarly denotes this preference.

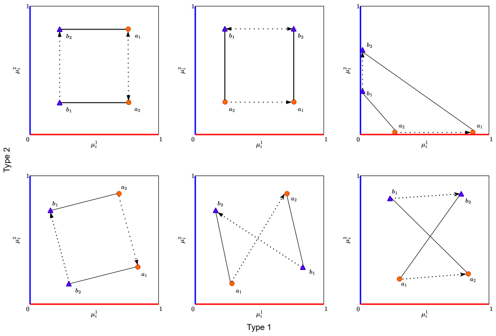

Figure 1: MP v.s. JP.

We refer to the above preference setting as *marginal preferences* (MP). To illustrate the distinction between marginal preferences and joint (couple) preferences (JP) (Che et al., 2019), we provide an example in Figure 1 involving two types of workers as an example. In the MP setting, preferences of each type of worker are independent of those of the other types. Here we use x-axis to represent the firm p_i's preference levels or called utility ($\mu_i^1 \in [0, 1]$) over type 1 workers (a_1, a_2), shown as red circles, and we use the y-axis to represent the firm p_i's preference levels ($\mu_i^2 \in [0, 1]$) over type 2 workers (b_1, b_2), shown as blue triangles. The first row of Figure 1 represents the MP setting, while the second row represents *the JP that the MP cannot cover*. In the MP setting, we show three possible patterns of combination matchings, $\{(a_1, b_2), (a_2, b_1)\}$ where two types of workers a and b are matched by firm, concatenated a solid line. Besides, a dashed arrow $(a_1 \rightarrow a_2)$ is used to represent the preference over matching, indicating that $a_1 <_i a_2$. Thus, the first row can be captured by the MP setting as the preference is consistent across the two types of workers. In contrast, the second row cannot be represented by the MP as it is not possible to compare the preference between (a_1, b_2) and (a_2, b_1) using $\{(a_1 > a_2), (b_2 < b_1)\}$. This MP setting is similar to the responsive preferences as defined in (Roth, 1985). Here we only consider the MP and the finding of efficient algorithm to solve the JP problem is notoriously difficult and unsolved (Che et al., 2019) and deserved more efforts.

(III) Matching Policy. At time t, for each firm p_i, $u_t^m(p_i) : \mathcal{N} \mapsto \mathcal{K}_m \cup \emptyset$ is a mapping function that satisfies $u_t^m(p_i) \in \mathcal{K}_m \cup \emptyset, \forall i \in [N]$, where $\emptyset$ represents a null set. At each time t, the platform assigns $m-$ type workers $u_t^m(p_i)$ for firm p_i. We define $u_t^m(p_i)$ as the function that maps each firm p_i in the set $\mathcal{N}$ to a set of $m-$ type workers $\mathcal{K}_m$ at time t. The assignments $u_t^m(p_i)$ for each firm are not only based on the firm's proposals (submitted preferences), but also on the *competing status* with other firms and workers' preferences. A centralized platform, such as LinkedIn or Amazon Mechanical Turks, coordinates this matching process in this competitive environment.

(IV) Stable Matching. The concept of *stability* is a widely used notion in the literature of stable matching, which refers to the property that no pair of agents (e.g., firms and workers) would mutually prefer each other over their current match (Gale and Shapley, 1962; Roth, 2008). This property is typically formalized as the absence of *blocking pairs* in the matching literature, which are pairs of agents that would both prefer to be matched with each other over their current match. The formal definition is illustrated as below.

Definition 1. *(Blocking pair). A matching u is blocked by a firm p_i if p_i prefers being single to being matched with $u(p_i)$, i.e. $p_i >_i u(p_i)$. A matching u is blocked by a pair of firm and worker (p_i, a_j) if they each prefer each other to the partner they receive at u, i.e. $a_j >_i u(p_i)$ and $p_i >_j u^{-1}(a_j)$.*

Roth (2008) stated if some preferences are not strict, arbitrarily breaking ties lets each agent fill out a strict preference list.

Definition 2. *(Stable Matching). A matching u is stable if it isn't blocked by any individual or pair of worker and firm.*

In this setting, however, each firm has a minimum quota vector $\mathbf{q}_i = [q_i^1, ..., q_i^M] \in \mathbb{R}^M$ for each type of worker to fill. Therefore, we define the concept of *stability* as the absence of "blocking pairs" across all types of workers and firms. Based on the definition of the stable matching, we also discussed the feasibility of the stable matching in the Appendix A. Here without loss of generality, we assume there exists the stable matching in the scheme of the complementary preference.

(V) Matching Reward. At time t, when firm p_i is matched with worker a_j^m, the firm receives a stochastic reward $y_{i,j}^m(t)$ which is assumed to be the *true matching reward* $\mu_{i,j}^m(t)$ plus a noise $\epsilon_{i,j}^m(t)$,

$$y_{i,j}^m(t) = \mu_{i,j}^m(t) + \epsilon_{i,j}^m(t), \forall i \in [N], \forall j \in [K_m], \forall m \in [M], \forall t \in [T], \tag{1}$$

where we assume that $\epsilon_{i,j}^m(t)$'s are independently drawn from a sub-Gaussian random variable with parameter σ. That is, for every $\alpha \in \mathbb{R}$, it is satisfied that $\mathbb{E}[\exp(\alpha \epsilon_{i,j}^m(t))] \leq \exp(\alpha^2 \sigma^2 / 2)$. The goal of the centralized platform is to design a learning algorithm that achieves stable matchings through learning the firms' preferences for multiple types of workers preciously from the previous matchings.

(VI) Regret. Based on model (1), we can observe a matching reward $\mathbf{y}_i^m(t) := \mathbf{y}_{i,u_t^m(p_i)}(t)$ at time t when firm p_i is matched with the assigned *m* $-$ type workers $u_t^m(p_i)$. The mean reward $\mu_{i,u_t^m(p_i)}(t)$ represents the noiseless reward (or mean utility) of firm p_i wrt its assigned matching *m*$-$type workers $\mathbf{u}_t^m(p_i)$ at time t. We define the cumulative *firm-optimal regret with m-type worker* for firm p_i as

$$R_i^m(T, \theta) := \sum_{t=1}^{T} \mu_{i,\overline{u}_i^m} - \sum_{t=1}^{T} \mu_{i,u_t^m(p_i)}(t), \tag{2}$$

where we denote θ as the sampled problem instance, and it is independently generated from a distribution Θ. This firm-optimal regret represents the difference between the capability of a policy $u_i^m := \{u_t^m(p_i)\}_{t=1}^T$ in hindsight and the optimal stable matching *oracle policy* $\overline{u}_i^m$. As each firm must recruit M types of workers with total quota Q_i, the *total cumulative firm-optimal stable regret* for firm p_i is defined as the sum of this difference over all types of workers, $R_i(T, \theta) := \mathbb{E}\left[\sum_{m=1}^M \sum_{t=1}^T \mu_{i,\overline{u}_i^m} - \sum_{m=1}^M \sum_{t=1}^T \mu_{i,u_t^m(p_i)}(t) | \theta\right]$. Finally, the *Bayesian total cumulative firm-optimal stable regret* for all firms is defined as the expected value of the total cumulative firm-optimal stable regret over all firms, $\mathfrak{R}(T) := \mathbb{E}_{\theta \in \Theta}\left[\sum_{i=1}^N R_i(T, \theta)\right]$. Our goal is to design an algorithm that minimizes this value over the time horizon T.

2.2 Challenges and Solutions

When preferences are unknown a priori in matching markets, the stability issue while satisfying complementary preferences and quota requirements is a challenging problem due to the interplay of multiple factors.

Challenge 1: How to design a stable matching algorithm for markets with complementary preferences? This is a prevalent issue in real-world applications such as hiring workers with complementary skills in hospitals and high-tech firms or admitting students with diverse backgrounds in college admissions. Despite its importance, no implementable algorithm is currently available to solve this challenge. In this paper, we propose a novel approach to resolving this issue by utilizing a *double matching* (Algorithm 3) to *marginalize* complementary preferences and achieve stability. Our algorithm can efficiently learn a stable matching solution using historical matching data, providing a practical solution to the problem of competing matching under complementary preferences.

Challenge 2: How to balance the exploration and exploitation to achieve the sublinear regret? The centralized platform must find a way to collect more firm-worker matching feedback while also achieving optimal matching at each time step. Compared to traditional matching algorithms, the CMCP is more challenging as it requires more time to balance this trade-off. In previous research, the classic UCB method could not achieve sublinear regret in some scenarios (Liu et al., 2020). We will also show an example in Section 3.2 to illustrate it. To overcome this challenge, we propose the use of TS algorithm which allows for random exploration and achieves sublinear regret.

Algorithm 1: Multi-agent Multi-type Thompson Sampling (MMTS)

```
Input :Time horizon $T$; firms' priors $(\boldsymbol{\alpha}_i^{m,0}, \boldsymbol{\beta}_i^{m,0}), \forall i \in [N], \forall m \in [M]$; workers' preference
       $\boldsymbol{\pi}^m, \forall m \in [M]$.
for $t \in \{1, ..., T\}$ do
    STEP 1: SAMPLING STAGE
        Get all firms' estimated rankings $\hat{\mathbf{r}}_i^m(t)$ and estimated mean reward $\hat{\boldsymbol{\mu}}_i^m(t)$ over all types
      of workers, $\forall i \in [N], m \in [M]$ from the Sampling stage in Algo 2.
    STEP 2: DOUBLE MATCHING STAGE
        Get the matching result $\mathbf{u}_t^m(p_i), \forall i \in [N], m \in [M]$ from the *double matching* in Algo 3.
    STEP 3: COLLECTING REWARDS STAGE
        Each firm receives its corresponding rewards from all types of workers $\mathbf{y}_i^m(t)$.
    STEP 4: UPDATING BELIEF STAGE
        Based on received rewards, firms update their posterior belief.
```

Challenge 3: How to solving CMCP with quota constraints in large markets? Unlike the classic DA algorithm (Gale and Shapley, 1962), our problem involves type-specific and total quota requirements for each firm. Can we find a stable matching algorithm that satisfies these constraints while also adapting to unknown preferences? Furthermore, can this algorithm be applied in large markets with efficiency? We address these challenges by proposing a novel algorithm, double matching, that effectively balances exploration and exploitation while can also be partially parallel implemented.

3 ALGORITHMS

In this section, we propose the Multi-agent Multi-type Thompson Sampling algorithm (MMTS), which aims to learn the true preferences of all firms over all types of workers, achieve stable matchings, and maximize the firms' Bayesian expected reward. We provide a detailed description of the algorithm and demonstrate its benefits of using TS. Besides, we also discuss the computational complexity of MMTS in Appendix B.

3.1 ALGORITHM DESCRIPTION

The MMTS in Algorithm 1 is composed of four stages, *sampling stage*, *double matching stage*, *collecting reward stage*, and *updating belief stage*. The common knowledge for centralized platform is the time horizon T, the number of participants, workers' preference $\{\boldsymbol{\pi}^m\}_{m=1}^M$, and firms' learning priors $\{(\boldsymbol{\alpha}_i^{m,0}, \boldsymbol{\beta}_i^{m,0})\}_{m=1}^M, \forall i \in [N]$. Then at each matching step t, MMTS iterates these four steps.

Step 1: Sampling Stage. For each firm p_i, it samples the estimated mean reward $\hat{\mu}_{i,j}^m(t)$ for $m-$ type worker a_j^m from a specific distribution $\mathcal{P}_j^m$ (e.g., Gaussian or beta distribution) with learned parameters $(\alpha_{i,j}^{m,t-1}, \beta_{i,j}^{m,t-1})$ from the previous time step $t-1$, which is $\hat{\mu}_{i,j}^m(t) \sim \mathcal{P}(\alpha_{i,j}^{m,t-1}, \beta_{i,j}^{m,t-1}), \forall i \in [N], \forall m \in [M], \forall j \in [\mathcal{K}_m]$. Besides, for the firm p_i, it sorts these type-specific workers based on the sampled mean reward $\{\hat{\mu}_{i,j}^m(t)\}_{j=1}^{K_m}$ in descending order and gets the estimated rank $\hat{\mathbf{r}}_i^m(t)$ for $m-$ type workers. After that, all firms submit their estimated ranks to the centralized platform. The above steps are shown in Algorithm 2.

Step 2: Double matching stage. With the shared estimated mean rewards $\hat{\boldsymbol{\mu}}(t) := \{\hat{\mu}_{i,j}^m(t)\}_{i,j,m}$ and estimated ranks $\hat{\mathbf{r}}(t) := \{\hat{\mathbf{r}}_i^m(t)\}_{i,m}$ from firms at time t, the double matching algorithm takes these ranks and quota constraints from firms as input and match firms and workers in two-stage matchings as shown in Algorithm 3. The goal of the first match is to allow all firms to satisfy their minimum type-specific quota q_i^m first. The second match is to fill the left-over positions $\tilde{Q}_i$ (defined below) for each firm and match firms and workers without type consideration. Before implementing the second match, we have to sanitize the status quo as a priori.

First Match: The platform implements the type-specific DA in Appendix Algorithm 4 given quota requirement $\{q_i^m\}_{m=1}^M, \forall i \in [N]$. The matching road map starts from matching all firms with type from 1 to M and returns the matching result $\{\tilde{u}_t^m(p_i)\}_{m\in[M]}$, which can be implemented in parallel.

```
Algorithm 2: Sampling Stage
Input :Time horizon $T$; firms' priors $(\boldsymbol{\alpha}_i^{m,0}, \boldsymbol{\beta}_i^{m,0}), \forall i \in [N], \forall m \in [M]$.
Sample: Sample mean reward $\hat{\mu}_{i,j}^m(t) \sim \mathcal{P}(\alpha_{i,j}^{m,t-1}, \beta_{i,j}^{m,t-1}), \forall i \in [N], \forall m \in [M], \forall j \in [\mathcal{K}_m]$.
Sort: Sort estimated mean rewards $\hat{\mu}_{i,j}^m(t)$ in descending order and get the estimated rank $\hat{\mathbf{r}}_i^m(t)$.
Output: The estimated rank $\hat{\mathbf{r}}_i^m(t)$ and the estimated mean rewards $\hat{\boldsymbol{\mu}}_i^m(t), \forall i \in [N], m \in [M]$.
```

```
Algorithm 3: Double Matching
Input :firms' estimated rank $\hat{\mathbf{r}}_i^m(t)$, estimated mean $\hat{\boldsymbol{\mu}}_i^m(t)$, type quota $q_i^m, \forall m \in [M], i \in [N]$
       and total quota $Q_i, \forall i \in [N]$; workers' preference $\{\boldsymbol{\pi}^m\}_{m \in [M]}$.
STEP 1: FIRST MATCH
    Submit all firms' estimated ranks $\hat{\mathbf{r}}_i^m(t)$ and all workers' preferences $\boldsymbol{\pi}^m$ to the platform.
    Run the firm choice DA in Algo 4 and return the matching $\tilde{u}_t^m(p_i)$ for firms over all types.
STEP 2: SANITIZE QUOTA
    Sanitize whether all firms' positions have been filled. For each company $p_i$, if
  $Q_i - \sum_{m=1}^M q_i^m > 0$, set the left quota as $\tilde{Q}_i \leftarrow Q_i - \sum_{m=1}^M q_i^m$ for firm $p_i$.
STEP 3: SECOND MATCH
if $\tilde{\mathbf{Q}} \neq 0$ then
    Submit left quota $\{\tilde{Q}_i\}_{i \in [N]}$, estimated means $\hat{\boldsymbol{\mu}}(t)$, and workers' preferences $\{\boldsymbol{\pi}^m\}_{m \in [M]}$
      to the centralized platform. Run the firm choice DA Algo 5 and return the matching $\breve{u}_t(p_i)$.
else
    Set the matching $\breve{u}_t(p_i) = \emptyset$.
Output :The matching $u_t^m(p_i) \leftarrow \text{Merge}(\tilde{u}_t^m(p_i), \breve{u}_t(p_i))$ for all firms.
```

Sanitize Quota: After Step 1's first match, the centralized platform will sanitize each firm's left-over quota $\tilde{Q}_i = Q_i - \sum_{m=1}^M q_i^m, \forall i \in [N]$. If there exists a firm $p_i, s.t., \tilde{Q}_i > 0$, then the platform will step into the second match. For those firms like p_i whose leftover quota is zero $\tilde{Q}_i = 0$, their matched workers will skip the second match.

Second Match: When rest firms and workers continue to join in the second match, the centralized platform implements the standard DA in Algorithm 5 without type consideration. That is, each firm will re-rank the rest M types of workers who do not have a match in the first match, and fill available vacant positions. It is worth noting that in Algorithm 5, each firm will not propose to the previous workers who rejected him/her and already matched worked in Step 1. Then firm p_i gets the corresponding matched workers $\breve{u}_t(p_i)$ in the second match. Finally, the centralized platform merges the first and second results to obtain a final matching for firm p_i with $m-$ type worker $\mathbf{u}_t^m(p_i) = \text{Merge}(\tilde{u}_t^m(p_i), \breve{u}_t(p_i)), \forall i \in [N], m \in [M]$ at time step t.

Step 3: Collecting Rewards Stage. When the platform broadcasts the matching result $\mathbf{u}_t^m(p_i)$ to all firms, each firm then receives its corresponding stochastic reward $\mathbf{y}_i^m(t), \forall i \in [N], m \in [M]$.

Step 4: Updating Belief Stage. After receiving these noisy rewards, firms update their belief (posterior) parameters as follows, $(\boldsymbol{\alpha}_i^{m,t}, \boldsymbol{\beta}_i^{m,t}) = \text{Update}(\boldsymbol{\alpha}_i^{m,t-1}, \boldsymbol{\beta}_i^{m,t-1}, \mathbf{y}_i^m(t)), \forall i \in [N], \forall m \in [M]$.

In summary, Algorithm 2 samples mean rewards and ranks based on historical matching data. Algorithm 3 computes the stable matching based on the estimated rewards and ranks from the sampling step. Step 3 (collecting rewards) and Step 4 (updating belief) update learning parameters based on received feedback from the assigned matching.

3.2 INCAPABLE EXPLORATION

We show why the TS has an advantage over the UCB style method in estimating the ranks of workers. We even find that centralized UCB does achieve linear firm-optimal stable regret in some cases and show it in Appendix C with detailed experimental setting and analysis.Why TS is capable of avoiding the curse of linear regret? By the property of sampling shown in Algorithm 2. Firm p_i's initial prior over worker a_i is a uniform random variable, and thus $r_j(t) > r_i(t)$ with probability $\hat{\mu}_j \approx \mu_j$, rather than *zero*! This differs from the UCB style method, which cannot update a_i's upper bound due to lacking exploration over a_i. The benefit of TS is that it can occasionally explore different ranking patterns, especially when there exists such a previous example. In Figure 2(a), we show a quick comparison of centralized UCB (Liu et al., 2020) in the settings shown above and MMTS when $M = 1, Q = 1, N = 3, K = 3$. The UCB method occurs a linear regret for firm 1 and firm 2. However, TS method suffers a sublinear regret in firm 1 and firm 2.

4 Properties of MMTS: Stability and Regret

In Section 4.1, we demonstrated the double matching technique providing the stability property for CMCP. Then we established the Bayesian regret upper bound for all firms when they follow the MMTS in Section 4.2. And we discussed the incentive-compatibility of the MMTS in Appendix G.

4.1 Stability

In the following theorem, we show the double matching technique can provide stable matching solution based on preferences from the firms' preferences over multiple types of workers provided by the MMTS and fixed and known preferences from workers.

Theorem 4.1. *Given two sides' strict preferences from firms and M types of workers. The double-matching procedure can provide a firm-optimal stable matching solution $\forall t \in [T]$.*

Proof. The detailed proof can be found in Appendix Section E. □

Remark. The sketch proof of the stability property of MMTS is two steps, naturally following the design of MMTS. The first match is conducted in parallel, and the output is stable and guaranteed by (Gale and Shapley, 1962). As the need of MMTS, before the second match, firms without leftover quotas ($\tilde{Q} = 0$) will quit the second round of matching, which will not affect the stability. After the quota sanitizing stage, firms and leftover workers will continue to join in the second matching stage, where firms do no need to consider the type of workers designed by double matching. And the standard DA algorithm will provide a stable result based on each firm's *sub-preference* list. The reason is that for firm p_i, all previous possible favorite workers have been proposed in the first match. If they are matched in the first match, they quit together, which won't affect the stability property; otherwise, the worker has a better candidate (firm) and has already rejected the firm p_i. So for each firm p_i, it only needs to consider a sub-preference list excluding the already matched workers in the first match and the proposed workers in the first match. It will provide a stable match in the second match and won't be affected by the first match. So the overall double matching is a stable algorithm.

4.2 Bayesian Regret Upper Bound

Next we provide the MMTS algorithm's Bayesian total cumulative firm-optimal regret upper bound.

Theorem 4.2. *Assume $K_{\max} = \max\{K_1, ..., K_M\}, K = \sum_{m=1}^{M} K_m$, with probability $1 - 1/QT$, when all firms follow the MMTS algorithm, firms together will suffer the Bayesian expected regret $\mathfrak{R}(T) \leq 8Q \log(QT)\sqrt{K_{\max}T} + NK/Q$.*

Proof. The detailed proof can be found in Appendix F. □

Remark. The derived Bayesian regret bound, which is dependent on the square root of the time horizon T and a logarithmic term, is nearly rate-optimal. Additionally, we examine the dependence of this regret bound on other key parameters. The first of which is a near-linear dependency on the total quota Q. Secondly, the regret bound is dependent only on the *square root* of the maximum worker $K_{\max}$ of one type, as opposed to the total number of workers, $\sum_{m=1}^{M} K_m$ in previous literature (Liu et al., 2020; Jagadeesan et al., 2021). This highlights the ability of our proposed algorithm, MMTS, to effectively capture the interactions of multiple types of matching in CMCP. The second term in the regret is a constant which is only dependent on constants N, K and the total quota Q. Notably, if we assume that each $q_i = 1$ and $Q_i = M$, then NK/Q will be reduced to $NK/(NM) = K/M$, which is an unavoidable regret term due to the exploration in bandits (Lattimore and Szepesvári, 2020). This also demonstrates that the Bayesian total cumulative firm-optimal exploration regret is only dependent on the *average* number of workers of each type available in the market, as opposed to the *total* number of workers or the maximum number of workers available of all types. Additionally, if one Q_i is dominant over other firms' Q_i, then the regret will mainly be determined by that dominant quota Q_i and $K_{\max}$, highlighting the inter-dependence of this complementary matching problem.

5 EXPERIMENTS

In this section, we present simulation results to demonstrate the effectiveness of MMTS in learning the unknown preferences of firms. The overall experiment setup can be found in Appendix I. In Section 5.1, we present two examples and analyze the underlying causes of the interesting phenomenon of negative regret. In Appendix I.2, we showcase the learning parameters from MMTS and provide insight into the reasons for non-optimal stable matchings. Additionally, we demonstrate the robustness of MMTS in large markets in Appendix I.3. All simulation results are run in 100 trials.

5.1 TWO EXAMPLES

Example 1. There are $N = 2$ firms, $M = 2$ types of workers, and there are $K_m = 5$ free workers for each type. The quota q_i^m for each type and each firm p_i is 2, and the total quota/capacity for each firm is $Q_i = 5$. The time horizon is $T = 2000$.

Preferences. True preferences from all types of workers to firms and from firms to different types of workers are all randomly generated. Workers to firms' preferences $\{\boldsymbol{\pi}^m\}_{m=1}^M$ are fixed and known. We use the data scientist (*D*/*DS*) and software developer engineer (*S*/*SDE*) as our example. The following are randomly generated true preferences for two-sided participants,

$$\begin{aligned}
&D_1 : p_1 \succ p_2, \quad D_2 : p_1 \succ p_2, \quad D_3 : p_2 \succ p_1, \quad D_4 : p_1 \succ p_2, \quad D_5 : p_2 \succ p_1,\\
&S_1 : p_1 \succ p_2, \quad S_2 : p_1 \succ p_2, \quad S_3 : p_2 \succ p_1, \quad S_4 : p_2 \succ p_1, \quad S_5 : p_1 \succ p_2,\\
&\pi_1^1 : D_4 \succ D_2 \succ D_3 \succ D_5 \succ D_1, \qquad \pi_1^2 : S_1 \succ S_4 \succ S_5 \succ S_2 \succ S_3,\\
&\pi_2^1 : D_2 \succ D_3 \succ D_1 \succ D_5 \succ D_4, \qquad \pi_2^2 : S_4 \succ S_2 \succ S_5 \succ S_1 \succ S_3.
\end{aligned} \tag{3}$$

The true matching reward of each worker for the firm is randomly generated from the uniform distribution $U([0, 1])$, and shown in Appendix Table 1. In addition, noisy reward $y_{i,j}^m(t)$ received (0 or 1) by each firm is generated by the Bernoulli distribution $y_{i,j}^m(t) \sim \text{Ber}(\mu_{i,j}^m(t))$, where $\mu_{i,j}^m(t)$ is the true matching reward at time t. If two sides' preferences are known, the firm optimal stable matching is $\bar{u}_1 = \{[D_2, D_4], [S_5, S_1, S_3]\}$, $\bar{u}_2 = \{[D_3, D_1, D_5], [S_4, S_2]\}$ by the double matching algorithm. However, if firms' preferences are unknown, MMTS can learn these unknown preferences and attain the optimal stable matching while achieving a sublinear regret for each firm.

MMTS Parameters. We set priors $\alpha_{i,j}^{m,0} = \beta_{i,j}^{m,0} = 0.1, \forall i \in [N], \forall j \in [K_m], \forall m \in [M]$ to avoid too strong impact of the prior information. Each firm follows the MMTS algorithm to propose multiple types of workers. The update formula for each firm p_i at time t of the m-type worker a_j^m is $\alpha_{i,j}^{m,t+1} = \alpha_{i,j}^{m,t} + 1$ if the worker a_j^m is matched with the firm p_i, that is $a_j^m \in \mathbf{u}_t^m(p_i)$, and the received reward for firm p_i is $y_{i,j}^m(t) = 1$; otherwise $\alpha_{i,j}^{m,t+1} = \alpha_{i,j}^{m,t}$; $\beta_{i,j}^{m,t+1} = \beta_{i,j}^{m,t} + 1$ if the worker a_j^m is matched with the firm p_i and the received reward for firm p_i is $y_{i,j}^m(t) = 0$, otherwise $\beta_{i,j}^{m,t+1} = \beta_{i,j}^{m,t}$. For other unmatched pairs (firm, $m-$ type worker), the parameters retain.

Results. In Figure 2(b), we find that firm 1, 2 achieve a total *negative* sublinear regret and a total *positive* sublinear regret separately (solid lines). However, we find that due to the incorrect rankings provided by firms, firm 1 benefits from this non-optimal matching result in terms of *negative* sublinear regret specifically for matching with type 1 workers (blue dashed line). More discussion about the negative regret phenomenon is available in Appendix I.1.

Example 2. We enlarge the market by expanding the DS market, particularly wanting to explore interactions between two types of workers. $N = 2$ firms, $M = 2$ types, $K_1 = 20$ (DS) and $K_2 = 6$ (SDE). The DS quota for two firms is $q_1^1 = q_2^1 = 1$ and the SDE quota for two firms is $q_1^2 = q_2^2 = 3$, and the total quota is $Q_i = 6$ for both firms. Preferences from firms to workers and workers to firms are randomly generated. Therefore, the matching result for each firm should consist of three workers for each type, and type II workers will be fully allocated in the first match, and the rest workers are all type II workers. All MMTS initial parameters are the same as in Example 1.

Results. In Figure 2(c), we show when excessive type II workers exist, and type I workers are just right. Both firms can achieve positive sublinear regret. We find that since type II worker $K_2 = q_1^2 + q_2^2 = 6$, which means in the first match stage, those type II workers are fully allocated into two firms. Thus, in the second match stage, the left quota would be all allocated to the type I workers

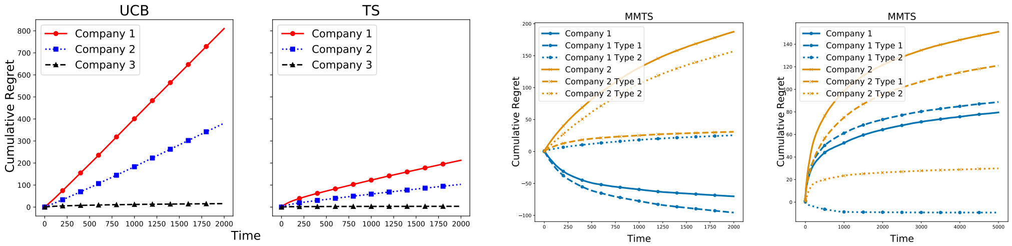

Figure 2: Left: A comparison of centralized UCB and TS. Right: firms and their sub-types regret for Example 1 and, firms and their sub-types regret for Example 2.

for two firms. Two dotted lines represent type II regret suffered by two firms. Both firms can quickly find the type II optimal matching since finding the optimal type II match just needs the first stage of the match. However, the type I workers' matching takes a longer time to find the optimal matching (take two stages), represented by dashed lines, and both are positive sublinear regret. Therefore, these two types of matching are fully independent, which is different from Example 1.

6 Related Works

We review multiple works in the literature, including matching while learning, multi-agent systems, assortment optimization, and matching markets. More can be found in Appendix J.

Matching while Learning. Liu et al. (2020) considers the multi-agent multi-armed competing problem in the centralized platform with explore-then-commit (ETC) and upper confidence bound (UCB) style algorithms where preferences from agents to arms are unknown and need to be learned through streaming interactive data. Jagadeesan et al. (2021) considers the two-sided matching problem where preferences from both sides are defined through dynamic utilities rather than fixed preferences and provide regret upper bounds over different contexts settings, and Min et al. (2022) apply it to the Markov matching market. Cen and Shah (2022) show that if there is transfer between agents, then the three desiderata (stability, low regret, and fairness) can be simultaneously achieved. Li et al. (2022) discuss the two-sided matching problem when the arm side has dynamic contextual information and preference is fixed from the arm side and propose a centralized contextual ETC algorithm to obtain the near-optimal regret bound. Besides, there are a plethora of works discussing the two-sided matching problem in the decentralized markets (Liu et al., 2021; Basu et al., 2021; Sankararaman et al., 2021; Dai and Jordan, 2021a;b; Dai et al., 2022). In particular, Dai and Jordan (2021b) study the college admission problem and provides an optimal strategy for agents, and shows its incentive-compatible property. Moreover, Jagadeesan et al. (2022) explores the phenomenon of the two-sided matching problem with two competing markets.

7 Conclusion and Future Work

In this paper, we proposed a new algorithm, MMTS to solve the CMCP. MMTS builds on the strengths of TS for exploration and employs a *double matching* method to find a stable solution. Through theoretical analysis, we show the effectiveness of the algorithm in achieving stability at every matching step, achieving a sublinear Bayesian regret over time, and exhibiting the IC property.

There are several directions for future research. One is to investigate more efficient exploration strategies to reduce the time required to learn the agents' unknown preferences. Another is to examine scenarios where agents have indifferent preferences, and explore the optimal strategy for breaking ties. Additionally, it is of interest to incorporate real-world constraints such as budget or physical locations into the matching process, which could be studied using techniques from constrained optimization. Moreover, it is interesting to incorporate side information, such as agents' background information, into the matching process. This can be approached using techniques from recommendation systems or other machine learning algorithms that incorporate side information. Finally, it would be interesting to extend the algorithm to handle time-varying matching markets where preferences and the number of agents may change over time.

REFERENCES

H. Abeledo and U. G. Rothblum. Paths to marriage stability. *Discrete applied mathematics*, 63(1): 1–12, 1995.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

A. Aouad and D. Saban. Online assortment optimization for two-sided matching platforms. *Management Science*, 2022.

I. Ashlagi, M. Braverman, and A. Hassidim. Matching with couples revisited. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 335–336, 2011.

I. Ashlagi, A. K. Krishnaswamy, R. Makhijani, D. Saban, and K. Shiragur. Assortment planning for two-sided sequential matching markets. *Operations Research*, 70(5):2784–2803, 2022.

P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

E. M. Azevedo and J. W. Hatfield. Existence of equilibrium in large matching markets with complementarities. *Available at SSRN 3268884*, 2018.

H. Aziz, J. Chen, S. Gaspers, and Z. Sun. Stability and pareto optimality in refugee allocation matchings. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 964–972, 2018.

S. Basu, K. A. Sankararaman, and A. Sankararaman. Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pages 705–715. PMLR, 2021.

P. Biró, D. F. Manlove, and I. McBride. The hospitals/residents problem with couples: Complexity and integer programming models. In *International Symposium on Experimental Algorithms*, pages 10–21. Springer, 2014.

N. Boehmer and K. Heeger. A fine-grained view on stable many-to-one matching problems with lower and upper quotas. *ACM Transactions on Economics and Computation*, 10(2):1–53, 2022.

S. H. Cen and D. Shah. Regret, stability & fairness in matching markets with bandit learners. In *International Conference on Artificial Intelligence and Statistics*, pages 8938–8968. PMLR, 2022.

Y.-K. Che, J. Kim, and F. Kojima. Stable matching in large economies. *Econometrica*, 87(1):65–110, 2019.

X. Dai and M. Jordan. Learning in multi-stage decentralized matching markets. *Advances in Neural Information Processing Systems*, 34:12798–12809, 2021a.

X. Dai and M. I. Jordan. Learning strategies in decentralized matching markets under uncertain preferences. *Journal of Machine Learning Research*, 22:260–1, 2021b.

X. Dai, Y. Qi, and M. I. Jordan. Incentive-aware recommender systems in two-sided markets. *arXiv preprint arXiv:2211.15381*, 2022.

L. E. Dubins and D. A. Freedman. Machiavelli and the gale-shapley algorithm. *The American Mathematical Monthly*, 88(7):485–494, 1981.

T. Fiez, B. Chasnov, and L. J. Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

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

D. González-Sánchez and O. Hernández-Lerma. *Discrete–time stochastic control and dynamic potential games: the Euler–Equation approach*. Springer Science & Business Media, 2013.

M. Greinecker and C. Kah. Pairwise stable matching in large economies. *Econometrica*, 89(6): 2929–2974, 2021.

J. Hadad and A. Teytelboym. Improving refugee resettlement: insights from market design. *Oxford Review of Economic Policy*, 38(3):434–448, 2022.

N. Immorlica, B. Lucier, V. Manshadi, and A. Wei. Designing approximately optimal search on matching platforms. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 632–633, 2021.

M. Jagadeesan, A. Wei, Y. Wang, M. Jordan, and J. Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

M. Jagadeesan, M. I. Jordan, and N. Haghtalab. Competition, alignment, and equilibria in digital marketplaces. *arXiv preprint arXiv:2208.14423*, 2022.

C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR, 2020.

B. Klaus and F. Klijn. Stable matchings and preferences of couples. *Journal of Economic Theory*, 121(1):75–106, 2005.

D. E. Knuth. Marriages stables. *Technical report*, 1976.

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

J. Komiyama, J. Honda, and H. Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161. PMLR, 2015.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Y. Li, C.-h. Wang, G. Cheng, and W. W. Sun. Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699*, 2022.

M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

M. L. Littman. Value-function reinforcement learning in markov games. *Cognitive systems research*, 2(1):55–66, 2001.

L. T. Liu, H. Mania, and M. Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

L. T. Liu, F. Ruan, H. Mania, and M. I. Jordan. Bandit learning in decentralized matching markets. *J. Mach. Learn. Res.*, 22:211–1, 2021.

D. F. Manlove, I. McBride, and J. Trimble. "almost-stable" matchings in the hospitals/residents problem with couples. *Constraints*, 22(1):50–72, 2017.

Y. Min, T. Wang, R. Xu, Z. Wang, M. I. Jordan, and Z. Yang. Learn to match with no regret: Reinforcement learning in markov matching markets. *arXiv preprint arXiv:2203.03684*, 2022.

T. Nguyen and R. Vohra. Near-feasible stable matchings with couples. *American Economic Review*, 108(11):3154–69, 2018.

T. Nguyen and R. Vohra. Complementarities and externalities. 2022.

J. Perolat, B. Piot, and O. Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pages 919–928. PMLR, 2018.

I. Rios, D. Saban, and F. Zheng. Improving match rates in dating markets through assortment optimization. *Manufacturing & Service Operations Management*, 2022.

A. E. Roth. The economics of matching: Stability and incentives. *Mathematics of operations research*, 7(4):617–628, 1982.

A. E. Roth. The college admissions problem is not equivalent to the marriage problem. *Journal of economic Theory*, 36(2):277–288, 1985.

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

A. E. Roth. Deferred acceptance algorithms: History, theory, practice, and open questions. *international Journal of game Theory*, 36(3):537–569, 2008.

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

D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

A. Sankararaman, S. Basu, and K. A. Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pages 1252–1260. PMLR, 2021.

C. Shi, R. Wan, G. Song, S. Luo, R. Song, and H. Zhu. A multi-agent reinforcement learning framework for off-policy evaluation in two-sided markets. *arXiv preprint arXiv:2202.10574*, 2022.

P. Shi. Optimal matchmaking strategy in two-sided marketplaces. *Management Science*, 2022.

T. Sönmez. Manipulation via capacities in two-sided matching markets. *Journal of Economic theory*, 77(1):197–204, 1997.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

K. Tomoeda. Finding a stable matching under type-specific minimum quotas. *Journal of Economic Theory*, 176:81–117, 2018.

C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar. Finite-sample analyses for fully decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:1812.02783*, 2018.

H. Zhong, Z. Yang, Z. Wang, and M. I. Jordan. Can reinforcement learning find stackelberg-nash equilibria in general-sum markov games with myopic followers? *arXiv preprint arXiv:2112.13521*, 2021.

M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

SUPPLEMENT TO TWO-SIDED COMPETING MATCHING MARKETS WITH COMPLEMENTARY PREFERENCES

This supplement is organized as follows. In Section A, we discuss the feasibility and its corresponding assumption of the stable matching. In Section B, we show the computational complexity of MMTS. In Section C, we exhibit why the centralized UCB suffers insufficient exploration. In Section D, we provide the Hoeffding concentration lemma. In Section E, we provide the stability property of MMTS. In Section F, we give the detailed proof of the regret upper bound of MMTS and decompose its proof into three parts, regret decomposition (F.1), bound for confidence width (F.2), and bad events' probabilities' upper bound (F.3). In Section G.1, we prove MMTS's strategy-proof property. Besides, as a reference, we append the DA with type and without type algorithms in Section H. In Section I, we provide details of experiments and the explanation of the negative regret, and also demonstrate the robustness of MMTS in large markets. Finally, in Section J, we provided additional related works.

A Feasibility of the Stable Matching

The feasibility solution is an interesting and well-discussed problem in the stable matching problem.

Assumption of the feasibility: In the finite market, it is the marginal preference assumption for the feasibility. But for the large market, it requires more assumptions such as the substitutability and indifferences, etc,. The difference between the infinite and finite (Azevedo and Hatfield, 2018; Greinecker and Kah, 2021) lies in matching problem and the techniques they use. In the infinite market, we assume that there is an uncountable number of agents on both sides of the market. This essentially means that the number of agents is so large that it can be treated as continuous, and you can't assign a specific numerical value to it. An example of an infinite market could be the matching of agents is extremely large and cannot be practically counted. In the finite market, the number of agents on both sides is limited and countable. You can assign a specific numerical value to the number of agents. An example could be the matching of agents where there is a definite small number of agents. However, such an exploration in the infinite market is beyond the scope of our current study.

In our case, if the complementary preference can be marginalized (or referred as the responsive preference (Roth, 1985), $(a_1, b_1) > (a_1, b_2)$ as long as $b_1 > b_2$, verse visa for $(a_1, b_1) > (a_2, b_1)$ as long as $a_1 > a_2$, which is at the top of Figure 1), then based on our proposed double matching algorithm and Theory 1, it exists such a stable matching solution. However, as discussed in the related works in Appendix J, if there exists couples in the preference list, which could potentially lead to an empty set of stable matchings.

Che et al. (2019) discussed that if there exists couples in the preference list in a infinite market (large) with a continuum of workers, provided that each firm's choice is convex and changes continuously as the set of available workers changes. They proved the existence and structure of stable matchings under preferences exhibiting substitutability and indifferences in a large market.

The difference between our result and (Che et al., 2019)'s result is in two ways: (1) we consider the finite market and they consider the infinite market. (2) we consider one side's preferences are unknown and (Che et al., 2019)'s both sides preferences are known. (3) Che et al. (2019) proved the existence of stable matching in the infinite market and no algorithm provided. However, in our paper, we provide the double matching algorithm to find it effectively.

B Complexity

Based on (Gale and Shapley, 1962; Knuth, 1997), the stable marriage problem's DA algorithm's worst total proposal number is $N^2 - 2N + 2 = \mathcal{O}(N^2)$ when the number of participants on both sides is equal ($N = K$). The computational complexity of the college admission matching problem with quota consideration is also $\mathcal{O}(NK)$. MMTS algorithm consists of two steps of matching. The computational complexity of the first step matching is $\mathcal{O}(\sum_{m=1}^{M} NK_m)$ if we virtually consider each type's matching process is organized in parallel. The second step's computation cost is also $\mathcal{O}(\sum_{m=1}^{M} NK_m)$. That is, in the first match, if all firms are matched with their best workers, this step meets the lower bound quota constraints. Then the second match will be reduced to the standard college admission problem without type consideration and the computational complexity is $\mathcal{O}(N\sum_{m=1}^{M} K_m)$. So the total computational complexity is still $\mathcal{O}(\sum_{m=1}^{M} NK_m)$, which is polynomial in the of firm (N) and the number of workers $\sum_{m=1}^{M} K_m$ in the market.

C Incapable Exploration

In this section, we show why the TS strategy has an advantage over the UCB style method in estimating the ranks of workers. We even find that centralized UCB does achieve linear firm-optimal stable regret in some cases. In the following example (Example 6 from (Liu et al., 2020)), we show the firm achieves linear optimal stable regret if follow the UCB algorithm.[3]

Let $\mathcal{N} = \{p_1, p_2, p_3\}$, $\mathcal{K}_m = \{a_1, a_2, a_3\}$, and $M = 1$, with true preferences given below:

$$\begin{aligned} p_1 &: a_1 \succ a_2 \succ a_3 & \qquad a_1 &: p_2 \succ p_3 \succ p_1 \\ p_2 &: a_2 \succ a_1 \succ a_3 & \qquad a_2 &: p_1 \succ p_2 \succ p_3 \\ p_3 &: a_3 \succ a_1 \succ a_2 & \qquad a_3 &: p_3 \succ p_1 \succ p_2 \end{aligned}$$

The firm optimal stable matching is $(p_1, a_1), (p_2, a_2), (p_3, a_3)$. However, due to incorrect ranking from firm p_3, $a_1 \succ a_3 \succ a_2$, and the output stable matching is $(p_1, a_2), (p_2, a_1), (p_3, a_3)$ based on the DA algorithm. In this case, p_3 will never have a chance to correct its mistake because p_3 will never be matched with a_1 again and cause the upper confidence bound for a_1 will never shrink and result in this rank $a_1 \succ a_3$. Thus, it causes that p_1 and p_2 suffer linear regret.

However, the TS is capable of avoiding this situation. By the property of sampling showed in Algorithm 2, firm p_1's initial prior over worker a_1 is a uniform random variable, and thus $r_3(t) > r_1(t)$ (if we omit a_2) with probability $\hat{\mu}_3 \approx \mu_3$, rather than *zero*! This differs from the UCB style method, which cannot update a_1's upper bound due to lacking exploration over a_1. The benefit of TS is that it can occasionally explore different ranking patterns, especially when there exists such a previous example.

In Figure 2(a), we show a quick comparison of centralized UCB (Liu et al., 2020) in the settings shown above and MMTS when $M = 1, Q = 1, N = 3, K = 3$. The UCB method occurs a linear regret in firm 1 and firm 2 and achieves a low matching rate (0.031)[4]. However, the TS method suffers a sublinear regret in firm 1 and firm 2 and achieves a high matching rate (0.741). All results are averaged over 100 trials. See Section C.1 for the experimental details.

C.1 Section 3.2 Example - Insufficient Exploration

We set the true matching reward for three firms to $(0.8, 0.4, 0.2)$, $(0.5, 0.7, 0.2)$, $(0.6, 0.3, 0.65)$. All preferences from companies over workers can be derived from the true matching reward. As we can view, company p_3 has a similar preference over a_1 (0.6) and a_3 (0.65). Thus, the small difference can lead the incapable exploration as described in Section 3.2 by the UCB algorithm.

D Hoeffding Lemma

Lemma D.1. *For any $\delta > 0$, with probability $1 - \delta$, the confidence width for a $m-$ type worker $a_j^m \in \mathcal{A}_{i,t}^m$ at time t is upper bounded by*

$$w_{i,\mathcal{F}_{i,t}^m}^m(a_j^m) \leq \min\left(2\sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}}, 1\right) \tag{C.1}$$

where $n_{i,j}^m(t)$ is the number of times that the pair (p_i, a_j^m) has been matched at the start of round t.

[3] Here we only consider one type of worker, and the firm's quota is one.

[4] We count 1 if the matching at time t is fully equal to the optimal match when two sides' preferences are known. Then we take an average over the time horizon T.

Proof. Let $\hat{\mu}_{i,j,t}^{m,LS} = \frac{\sum_{s=1}^{t} \mathbf{1}(a_j^m \in \mathcal{A}_{i,s}^m) y_{i,j}^m(s)}{n_{i,j}^m(t)}$ denote the empirical mean reward from matching firm p_i and $m-$ type worker a_j^m up to time t. Define upper and lower confidence bounds as follows:

$$U_{i,t}^m(a_j^m) = \min\left\{\hat{\mu}_{i,j,t}^{m,LS} + \sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}}, 1\right\}, L_{i,t}^m(a_j^m) = \max\left\{\hat{\mu}_{i,j,t}^{m,LS} - \sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}}, 0\right\}. \quad \text{(C.2)}$$

The the confidence width is upper bounded by $\min\left(2\sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}}, 1\right)$. □

E PROOF OF THE STABILITY OF MMTS

Proof. We shall prove existence by giving an iterative procedure to find a stable matching.

Part I To start, in the *first match* loop, based on the double matching procedure, we can discuss M types of matching in parallel. So we will only discuss the path for seeking the type-m company-worker stable matching.

Suppose firm p_i has q_i^m quota for m-type workers. We replace each firm p_i by q_i^m copies of p_i denoted by $\{p_{i,1}, p_{i,2}, ..., p_{i,q_i^m}\}$. Each of these $p_{i,h}$ has preferences identical with those of p_i but with a quota of 1. Further, each m-type worker who has p_i on his/her preference list now replace p_i by the set $\{p_{i,1}, p_{i,2}, ..., p_{i,q_i^m}\}$ in that order of preference. It is now easy to verify that the stable matchings for the firm m-type worker matching problem are in natural one-to-one correspondence with the stable matchings of this modified version problem. Then in the following, we only need to prove that stable matching exists in this transformed problem where each firm has quota 1, which is the standard stable marriage problem (Gale and Shapley, 1962). The existence of stable matching has been given in (Gale and Shapley, 1962). Here we reiterate it to help us to find the stable matching in the *second match*.

Let each firm propose to his favorite m-type worker. Each worker who receives more than one offer rejects all but her favorite from among those who have proposed to her. However, the worker does not fully accept the firm, but keeps the firm on a string to allow for the possibility that some better firm come along later.

Now we are in the second stage. Those firms who were rejected in the first stage propose to their second choices. Each m-type worker receiving offers chooses her favorite from the group of new firms and the firm on her string, if any. The worker rejects all the rest and again keeps the favorite in suspense. We proceed in the same manner. Those firms who are rejected at the second stage propose to their next choices, and the m-type workers again reject all but the best offer they have had so far.

Eventually, every m-type worker will have rejected a proposal, for as long as any worker has not been proposed to there will be rejections and new offers[5], but since no firm can propose the same m-type worker more than once, every worker is sure to get a proposal in due time. As soon as the last worker gets her offer, the “recruiting" is declared over, and each m-type worker is now required to accept the firm on her string.

We asset that this set of matching is stable. Suppose firm p_i and m-type worker a_j are not matched to each other but firm p_i prefers a_j to his current matching m-type worker $a_{j'}$. Then p_i must have proposed to a_j at some stage (since the proposal is ordered by the preference list) and subsequently been rejected in favor of some firm $p_{i'}$ that a_j liked better. It is clear that a_j must prefer her current matching firm $p_{i'}$ and there is no instability/blocking pair.

Thus, each m-type firm-worker matching established on the first match is stable. Then each firm p_i's matching object in the first match with quota q_i^m can be recovered as grouping all matching objects of firm $\{p_{i,h}\}_{h=1}^{q_i^m}$.

Part II To start the second match, we first check the left quota $\tilde{Q}_i$ for each firm. If the left quota is zero for firm p_i, then firm p_i and its matching workers will quit the matching market and get its stable matching object. Otherwise, the left firm will continue to participate in the second match.

[5] Here we assume the number of firms is less than or equal to the number workers, and those workers unmatched finally will be matched to themselves and assume their matching object is on the firm side.

In the second match, preferences from firms to workers are un-categorized. Based on line 19 in Algorithm 3, all types of workers will be ranked to fill the left quota. Thus, it reduces to the problem in part I, and the result matching in the second match is also stable. What is left to prove is that the overall double matching algorithm can provide stable matching. In the second match, each firm proposes to workers in his left concatenate ordered preference list, and all previous workers not in the second match preference list have already been matched or rejected. So it cannot form a blocking pair between the firm p_i with leftover workers. □

F MMTS Regret Upper Bound

F.1 Regret Decomposition

In this part, we provide the road map of the regret decomposition and key steps to prove Theorem 4.2. First, we define the history for firm p_i up to time t of type m as $H^m_{i,t} := \{\mathcal{A}^m_{i,1}, \mathbf{y}^m_{i,\mathcal{A}^m_{i,1}}(1), \mathcal{A}^m_{i,2}, \mathbf{y}^m_{i,\mathcal{A}^m_{i,2}}(2), ..., \mathcal{A}^m_{i,t-1}, \mathbf{y}^m_{i,\mathcal{A}^m_{i,t-1}}(t-1)\}$, composed by actions (matched workers) and rewards, where $\mathcal{A}^m_{i,t} := \mathbf{u}^m_t(p_i)$ is a set of workers (based on quota requirement q^m_i and Q_i) belong to m-type which is matched with firm p_i at time t, $\mathbf{y}^m_{i,\mathcal{A}^m_{i,t-1}}(t-1)$ are realized rewards when firm p_i matched with $m -$ type workers $\mathcal{A}^m_{i,t}$. Define $\widetilde{H}_{i,t} := \{H^1_{i,t}, H^2_{i,t}, ..., H^M_{i,t}\}$ as the aggregated interaction history between firm p_i and all types of workers up to time t.

Next, we define the *good event* for firm p_i when matching with $m -$ type worker at time t and the true mean matching reward falls in the uncertainty set as $E^m_{i,t} = \{\boldsymbol{\mu}^m_{i,\mathcal{A}^m_{i,t}} \in \mathcal{F}^m_{i,t}\}$, where $\boldsymbol{\mu}^m_{i,\mathcal{A}^m_{i,t}}$ is the true mean reward vector of actually pulled arms (matched with $m -$ type workers) at time t for firm p_i, and $\mathcal{F}^m_{i,t}$ is the uncertainty set for $m -$ type worker at time t for firm p_i. Similarly, the good event for firm p_i when matching with all types of workers at time t is $E_{i,t} = \bigcap_{m=1}^M E^m_{i,t}$, over all firms $E_t = \bigcap_{i=1}^N E_{i,t}$. And the corresponding *bad event* is defined as $\overline{E}^m_{i,t}, \overline{E}_{i,t}, \overline{E}_t$ respectively. That represents the true mean vector/tensor reward of the pulled arms is not in the uncertainty set.

Lemma F.1. *Fix any sequence $\{\widetilde{\mathcal{F}}_{i,t} : i \in [N], t \in \mathbb{N}\}$, where $\widetilde{\mathcal{F}}_{i,t} \subset \mathcal{F}$ is measurable with respect to $\sigma(\widetilde{H}_{i,t})$. Then for any $T \in \mathbb{N}$, with probability 1,*

$$\mathfrak{R}(T) \leq \mathbb{E} \sum_{t=1}^{T} \left[\sum_{i=1}^{N} \sum_{m=1}^{M} \widetilde{W}^m_{i,\mathcal{F}^m_{i,t}}(\mathcal{A}^m_{i,t}) + C\mathbf{1}(\overline{E}_t) \right] \tag{C.3}$$

where $\widetilde{W}^m_{i,\widetilde{\mathcal{F}}^m_{i,t}}(\cdot) = \sum_{a^m_j \in \mathcal{A}^m_{i,t}} w^m_{i,\widetilde{\mathcal{F}}^m_{i,t}}(a^m_j)$ represents the sum of the element-wise value of uncertainty width at m $-$ type worker a^m_j. The uncertainty width $w^m_{i,\widetilde{\mathcal{F}}^m_{i,t}}(a^m_j) = \sup_{\bar{\mu}^m_i, \underline{\mu}^m_i \in \widetilde{\mathcal{F}}^m_{i,t}} (\bar{\mu}^m_i(a^m_j) - \underline{\mu}^m_i(a^m_j))$ is a worst-case measure of the uncertain about the mean reward of m $-$ type worker a^m_j. Here C is a constant less than 1.

Proof. The key step of regret decomposition is to split the instantaneous regret by firms, types, and quotas. Then we categorize regret by the happening of good events and bad events. The good events' regret is measured by the uncertainty width, and the bad events' regret is measured by the probability of happening it.

To reduce notation, define element-wise upper and lower bounds $U^m_{i,t}(a) = \sup\{\mu^m_i(a) : \mu^m_i \in \mathcal{F}^m_{i,t}, a \in \mathcal{K}_m\}$ and $L^m_{i,t}(a) = \inf\{\mu^m_i(a) : \mu^m_i \in \mathcal{F}^m_{i,t}, a \in \mathcal{K}_m\}$, where μ^m_i is the mean reward function $\mu^m_i \in \mathcal{F}^m_{i,t} : \mathbb{R} \mapsto \mathbb{R}, \forall i \in [N], \forall m \in [M]$. Whenever $\mu^m_{i,\widetilde{\mathcal{A}}^m_i} \in \mathcal{F}^m_{i,t}$, the bounds $L^m_{i,t}(a) \leq \mu^m_{i,\widetilde{\mathcal{A}}^m_i}(a) \leq U^m_{i,t}(a)$ hold for all types of workers. Here we define $\mathcal{A}^m_{i,t} = \mathbf{u}^m_i(t)$ as the matched $m -$ type workers for firm p_i at time t and $\mathcal{A}^{m,*}_{i,t} = \overline{\mathbf{u}}^m_i(t)$ as the firm p_i's optimal stable matching result of $m -$ type workers at time t. Since the firm-optimal stable matching result is fixed, given both sides' preferences, we can omit time t here. The firm-optimal stable matching result set is also denoted as $\mathcal{A}^{m,*}_i = \mathcal{A}^{m,*}_{i,t}$.

As for type-m workers' matching for the firm p_i at time t, the instantaneous regret with a given instance θ can be implied as follows, here for simplicity, we omit the instance conditional notation

$$\begin{aligned}\mathcal{I}_{i,t}^m = \mu_i^m(\mathcal{A}_i^{m,*}) - \mu_i^m(\mathcal{A}_{i,t}^m) &\leq \sum_{a \in \mathcal{A}_i^{m,*}} U_{i,t}^m(a) - \sum_{a \in \mathcal{A}_{i,t}^m} L_{i,t}^m(a) + C\mathbf{1}(\boldsymbol{\mu}_{i,\widetilde{\mathcal{A}}_i}^m \notin \mathcal{F}_{i,t}^m) \\ &= \widetilde{U}_{i,t}^m(\mathcal{A}_i^{m,*}) - \widetilde{L}_{i,t}^m(\mathcal{A}_{i,t}^m) + C\mathbf{1}(\boldsymbol{\mu}_{i,\widetilde{\mathcal{A}}_i}^m \notin \mathcal{F}_{i,t}^m) \\ &= \widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\mathcal{A}_{i,t}^m) + [\widetilde{U}_{i,t}^m(\mathcal{A}_i^{m,*}) - \widetilde{U}_{i,t}^m(\mathcal{A}_{i,t}^m)] + C\mathbf{1}(\boldsymbol{\mu}_{i,\widetilde{\mathcal{A}}_i}^m \notin \mathcal{F}_{i,t}^m),\end{aligned} \tag{C.4}$$

where $C \leq 1$ is a constant, and we let $\widetilde{U}_{i,t}^m(\cdot) = \sum_a U_{i,t}^m(a)$ and $\widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\cdot) = \sum_a w_{i,\mathcal{F}_t}^m(a)$ represent the sum of the element-wise value of $U_{i,t}^m(\cdot), w_{i,\mathcal{F}_{i,t}}^m(\cdot)$, respectively. Define the good event for firm p_i, matching with $m -$ type worker at time t is $E_{i,t}^m = \{\boldsymbol{\mu}_{i,\widetilde{\mathcal{A}}_i}^m \in \mathcal{F}_{i,t}^m\}$, over all types $E_{i,t} = \bigcap_{m=1}^M E_{i,t}^m$, over all firms $E_t = \bigcap_{i=1}^N E_{i,t}$. And the corresponding bad event is defined as $\overline{E}_{i,t}^m, \overline{E}_{i,t}, \overline{E}_t$ respectively.

Now consider Eq. (C.3), summing over the previous equation over time t, firms p_i, and workers' type m, we get

$$\begin{aligned}\mathfrak{R}(T) &\leq \mathbb{E}\sum_{i=1}^N \sum_{t=1}^T \sum_{m=1}^M [\widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\mathcal{A}_{i,t}^m) + C\mathbf{1}(\overline{E}_t)] + \sum_{i=1}^N \mathbb{E}M_{i,T} \\ &= \mathbb{E}\sum_{t=1}^T [C\mathbf{1}(\overline{E}_t) + \sum_{i=1}^N \sum_{m=1}^M \widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\mathcal{A}_{i,t}^m)] + \sum_{i=1}^N \mathbb{E}M_{i,T}\end{aligned} \tag{C.5}$$

where $M_{i,T} = \sum_{t=1}^T \sum_{m=1}^M [\widetilde{U}_{i,t}^m(\mathcal{A}_i^{m,*}) - \widetilde{U}_{i,t}^m(\mathcal{A}_{i,t}^m)]$. Now by the definition of TS, $\mathbb{P}_m(\mathcal{A}_{i,t}^m \in \cdot|H_{i,t}^m) = \mathbb{P}_m(\mathcal{A}_i^{m,*} \in \cdot|H_{i,t}^m)$ for all types, where $\mathbb{P}_m(\cdot|H_{i,t}^m)$ represents this probability is conditional on history $H_{i,t}^m$ and the selected action (worker) belongs in m-type workers for firm p_i. That is $\mathcal{A}_{i,t}^m$ and $\mathcal{A}_i^{m,*}$ within type-m is identically distributed under the posterior. Besides, since the confidence set $\mathcal{F}_{i,t}^m$ is $\sigma(H_{i,t}^m)$-measurable, so is the induced upper confidence bound $U_{i,t}^m(\cdot)$. This implies $\mathbb{E}_m[U_{i,t}^m(\mathcal{A}_{i,t}^m)|H_t^m] = \mathbb{E}_m[U_{i,t}^m(\mathcal{A}_i^{m,*})|H_t^m]$, and there for $\mathbb{E}[M_{i,T}] = 0$ and $\sum_{i=1}^N \mathbb{E}M_{i,T} = 0$. Then we can obtain the desired result. □

F.2 Uncertainty Widths

In this part, we provide the upper bound of the accumulated uncertainty widths over all types of workers and all firms, which is the first part in Eq. (C.3).

Lemma F.2. *If $(\beta_{i,j,t}^m \geq 0|t \in \mathbb{N})$ is a non-decreasing sequence and $\mathcal{F}_{i,j,t}^m := \{\mu_{i,j}^m \in \mathcal{F}_{i,j}^m : \left\|\mu_{i,j}^m - \hat{\mu}_{i,j,t}^{m,LS}\right\|_1 \leq \sqrt{\beta_{i,j,t}^m}\}$, then with probability 1,*

$$\sum_{t=1}^T \sum_{i=1}^N \sum_{m=1}^M \widetilde{W}_{i,\mathcal{F}_{i,t}^m}^m(\mathcal{A}_{i,t}^m) \leq 8Q\log(QT)\sqrt{K_{\max}T}.$$

The proof of this lemma builds upon Lemma F.3, which establishes the number of instances where the widths of uncertainty sets for a chosen set of $m -$ type workers $\mathcal{A}_{i,t}^m$ greater than ϵ. We show that this number is determined by the *Eluder dimension* (Russo and Van Roy, 2014).

Proof. By Lemma F.1, the instantaneous regret $\mathcal{I}_t$ over all firms and all types, can be decomposed by types and by firms and shown as

$$\begin{aligned}\mathcal{I}_t &= \sum_{m=1}^{M} \mathcal{I}_t^m = \sum_{i=1}^{N}\sum_{m=1}^{M} \mathcal{I}_{i,t}^m \\ &\leq \sum_{i=1}^{N}\sum_{m=1}^{M} \widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\mathcal{A}_{i,t}^m), \quad \text{if } E_t \text{ holds.} \qquad \text{(C.6)} \\ &\leq 2 \sum_{i\in[N],m\in[M],a_j^m\in\mathcal{K}_m} \sqrt{\frac{\log(\sum_{i=1}^{N} Q_i T)}{n_{i,j}^m(t)}}, \quad \text{with prob } 1-\delta\end{aligned}$$

where the first inequality is based on Lemma F.1 and if E_t holds for $t \in \mathbb{N}, m \in M, i \in [N]$, $n_{i,j}^m(t)$ is the number of times that the pair (p_i, a_j^m) has been matched at the start of round t. The second inequality is constructed from a union concentration inequality based on Lemma D.1, and we set $\delta = 2/\sum_{i=1} Q_i T$. We denote $z_{i,j}^m(t) = \frac{1}{\sqrt{n_{i,j}^m(t)}}$ as the size of the scaled confidence set (without the log factor) for the pair (p_i, a_j^m) at the time t.

At each time step t, let's consider the list consisting of $z_{i,j}^m(t)$ and reorder the overall list consisting of concatenating all those scaled confidence sets over all rounds and all types in decreasing order. Then we obtain a list $\tilde{z}_1 \geq \tilde{z}_2 \geq ..., \geq \tilde{z}_L$, where $L = \sum_{t=1}^{T}\sum_{i=1}^{N} Q_i = T\sum_{i=1}^{N} Q_i$. We reorganize the Eq. (C.6) to get

$$\sum_{t=1}^{T} \mathcal{I}_t \leq \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{i=1}^{N} \widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\mathcal{A}_{i,t}^m) \leq 2\log(\sum_{i=1}^{N} Q_i T)\sum_{l=1}^{L} \tilde{z}_l. \qquad \text{(C.7)}$$

By Lemma F.3, the number of rounds that a pair of a firm and any $m-$ type worker can have it confidence set have size at least $\tilde{z}_l$ is upper bounded by $(1+\frac{4}{\tilde{z}_l^2})K_m$ when we set $\epsilon = \tilde{z}_l$ and know $\beta_{i,j,t}^m \leq 1$. Thus, the total number of times that any confidence set can have size at least $\tilde{z}_l$ is upper bounded by $\left(1+\frac{4}{\tilde{z}_l^2}\right)\sum_{i=1}^{N}\sum_{m=1}^{M} |\mathcal{A}_{i,t}^m| K_m$. To determine the minimum condition for $\tilde{z}_l$, which is equivalent to determine the maximum of l, we have $l \leq \left(1+\frac{4}{\tilde{z}_l^2}\right)\sum_{i=1}^{N}\sum_{m=1}^{M} |\mathcal{A}_{i,t}^m| K_m$. So we claim that

$$\tilde{z}_l \leq \min\left(1, \frac{2}{\sqrt{\frac{l}{\sum_{i=1}^{N}\sum_{m=1}^{M}|\mathcal{A}_{i,t}^m|K_m} - 1}}\right) \leq \min\left(1, \frac{2}{\sqrt{\frac{l}{\sum_{i=1}^{N} Q_i K_{\max}} - 1}}\right), \qquad \text{(C.8)}$$

where the second inequality above is by $\sum_{i=1}^{N}\sum_{m=1}^{M} |\mathcal{A}_{i,t}^m| K_m \leq K_{\max}\sum_{i=1}^{N}\sum_{m=1}^{M} |\mathcal{A}_{i,t}^m| \leq K_{\max}\sum_{i=1}^{N} Q_i = QK_{\max}$ and $K_{\max} = \max\{K_1, ..., K_M\}, Q = \sum_{i=1}^{N} Q_i$. Putting all these together, we have

$$\begin{aligned}2\log(\sum_{i=1}^{N} Q_i T)\sum_{l=1}^{L}\tilde{z}_l &\leq 2\log(QT)\sum_{l=1}^{L}\min(1, \frac{2}{\sqrt{\frac{l}{QK_{\max}} - 1}}) \\ &= 4\log(QT)\sum_{l=1}^{QT}\frac{1}{\sqrt{\frac{l}{QK_{\max}} - 1}} \qquad \text{(C.9)} \\ &\leq 8\log(QT)\sqrt{QK_{\max}}\sqrt{QT}\end{aligned}$$

where the last inequality is by intergral inequality

$$\sum_{l=1}^{QT}\frac{1}{\sqrt{\frac{l}{QK_{\max}} - 1}} \leq \sqrt{QK_{\max}}\sum_{l=1}^{QT}\frac{1}{\sqrt{l}} \leq \sqrt{QK_{\max}}\int_{x=0}^{QT}\frac{1}{\sqrt{x}}dx = 2\sqrt{QK_{\max}}\sqrt{QT}.$$

Based on Eq. (C.7) and the above result, we can get the regret

$$\sum_{t=1}^{T}\mathcal{I}_t \leq 8Q\log(QT)\sqrt{K_{\max}T}, \qquad \text{(C.10)}$$

if E_t holds. □

Lemma F.3. *If $(\beta^m_{i,j,t} \geq 0 | t \in \mathbb{N})$ is a nondecreasing sequence for $i \in [N], a^m_j \in \mathcal{K}_m, m \in [M]$ and $\mathcal{F}^m_{i,j,t} := \{\mu^m_{i,j} \in \mathcal{F}^m_{i,j} : \left\|\mu^m_{i,j} - \hat{\mu}^{m,LS}_{i,j,t}\right\|_1 \leq \sqrt{\beta^m_{i,j,t}}\}$, for all $T \in \mathbb{N}$ and $\epsilon > 0$, then*

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{a_j^m \in \mathcal{A}^m_{i,t}} \mathbf{1}\big(w^m_{i,\mathcal{F}^m_{i,t}}(a^m_j) > \epsilon\big) \leq \big(\frac{4\widetilde{\beta}_{i,T}}{\epsilon^2} + 1\big) \sum_{m=1}^{M} |\mathcal{A}^m_{i,t}| K_m.$$

Here $\hat{\mu}^{m,LS}_{i,j,t} = \frac{\sum_{s=1}^{t} \mathbf{1}(a^m_j \in \mathcal{A}^m_{i,s}) y^m_{i,j}(s)}{n^m_{i,j}(t)}$ is the estimated average reward for m − type worker a^m_j from the view point of firm p_i at time t, and $n^m_{i,j}(t)$ is the number of matched times up to time t of firm p_i with m − type worker a^m_j. Besides, we define $\widetilde{\beta}_{i,T} = \max_{a^m_j \in \mathcal{K}_m, m \in [M]} \beta^m_{i,j,T}$ as the maximum uncertainty bound over all types of workers at time T for firm p_i.

The proof of this result is based on techniques from (Russo and Van Roy, 2013; 2014). This result demonstrates that the upper bound of the number of times the widths of uncertainty sets exceeds ϵ is dependent on the error $\mathcal{O}(\epsilon^{-2})$ and linearly proportional to the product of the number of *m* − type worker and the type quota size q^m_i.

Proof. Based on the Proposition 3 from (Russo and Van Roy, 2013), we can use the *eluder dimension* $dim_E(\mathcal{F}^m_i, \epsilon)$ to bound the number of times the widths of confidence intervals for a selection of set of *m* − type workers $\mathcal{A}^m_{i,t}$ greater than ϵ.

$$\begin{aligned}\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{a_j^m \in \mathcal{A}^m_{i,t}} \mathbf{1}\bigg(w^m_{i,\mathcal{F}^m_{i,t}}(a^m_j) > \epsilon\bigg) &\leq \sum_{m=1}^{M}\sum_{a_j^m \in \mathcal{A}^m_{i,t}} \bigg(\frac{4\beta^m_{i,j,T}}{\epsilon^2} + 1\bigg)\mathrm{dim}_E(\mathcal{F}^m_i, \epsilon) \\ &\leq \left(\frac{4 \max\limits_{a^m_j \in \mathcal{K}_m, m \in [M]} \beta^m_{i,j,T}}{\epsilon^2} + 1\right) \sum_{m=1}^{M} |\mathcal{A}^m_{i,t}| \mathrm{dim}_E(\mathcal{F}^m_i, \epsilon),\end{aligned} \tag{C.11}$$

where the eluder dimension of a multi-arm bandit problem is the number of arms, we get

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{a_j^m \in \mathcal{A}^m_{i,t}} \mathbf{1}\bigg(w^m_{i,\mathcal{F}_t}(a^m_j) > \epsilon\bigg) \leq \bigg(\frac{4\widetilde{\beta}_{i,T}}{\epsilon^2} + 1\bigg) \sum_{m=1}^{M} |\mathcal{A}^m_{i,t}| K_m \leq \bigg(\frac{4\widetilde{\beta}_{i,T}}{\epsilon^2} + 1\bigg) Q_i K_{\max} \tag{C.12}$$

where $\widetilde{\beta}_{i,T} = \max_{a^m_j \in \mathcal{K}_m, m \in [M]} \beta^m_{i,j,T}$. Besides, we know that $Q_i = \sum_{m=1}^{M} |\mathcal{A}^m_{i,t}|$ and define $K_{\max} = \max_{m \in [M]} K_m$, so we can get the second inequality. □

F.3 Bad Event Upper Bound

In this part, we provide an upper bound of the second part of Eq. (C.3). The regret caused by the happening of the bad event at each time step is quantified by the following lemma.

Lemma F.4. *If $\mathcal{F}^m_{i,j,t} := \{\mu^m_{i,j} \in \mathcal{F}^m_{i,j} : \left\|\mu^m_{i,j} - \hat{\mu}^{m,LS}_{i,j,t}\right\|_1 \leq \sqrt{\beta^m_{i,j,t}}\}$ holds with probability $1 - \delta$, then the bad event $\overline{E}_t$ happening's probability is upper bounded by $\mathbb{E}\mathbf{1}(\overline{E}_t) \leq NK\delta$. In particular, if $\delta = 1/QT$, the accumulated bad events' probability is upper bounded by $\sum_{t=1}^{T} \mathbb{E}\mathbf{1}(\overline{E}_t) \leq NK/Q$.*

To bound the probability of bad events, we use a union bound to obtain the desired result. Specifically, if $Q_i = 1$, which means each firm has a total quota of 1 and only considers one type of worker, then $\sum_{t=1}^{T} \mathbb{E}\mathbf{1}(\overline{E}_t) \leq NK/(N \times 1) = K$. This shows that each firm needs to explore a single type of worker, and the worst total regret is less than K. If $Q_i = 1, M = 1$, which means all firms have the same recruiting requirements, the result reduces to the general competitive matching scenario, and the worst regret is the number of workers of type K_M in the market.

Proof. If E_t does not hold, the probability of the true matching reward is not in the confidence interval we constructed is upper bounded by

$$
\begin{aligned}
\mathbb{E}\mathbf{1}(\overline{E}_t) = \mathbb{P}(\overline{E}_t) &= \mathbb{P}\Bigg(\Big(\bigcap_{i\in[N]} \bigcap_{m\in[M]} \bigcap_{a_j^m\in\mathcal{K}_m} \{\mu_{i,j}^m \in \mathcal{F}_{i,j,t}^m\} \Big)^c \Bigg) \\
&= \mathbb{P}\Bigg(\bigcup_{i\in[N]} \bigcup_{a_j^m\in\mathcal{K}_m} \bigcup_{m\in[M]} \{\mu_{i,j}^m \notin \mathcal{F}_{i,j,t}^m\} \Bigg) \\
&= \mathbb{P}\Bigg(\bigcup_{i\in[N]} \bigcup_{a_j^m\in\mathcal{K}_m} \bigcup_{m\in[M]} \left\{ \left\| \mu_{i,j}^m - \hat{\mu}_{i,j,t}^{m,LS} \right\|_{2,E_t} \geq \sqrt{\beta_{i,j,t}^m} \right\} \Bigg) \qquad \text{(C.13)} \\
&= \mathbb{P}\Bigg(\bigcup_{i\in[N]} \bigcup_{a_j^m\in\mathcal{K}_m} \bigcup_{m\in[M]} \left\{ \left\| \mu_{i,j}^m - \hat{\mu}_{i,j,t}^{m,LS} \right\|_1 \geq \sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}} \right\} \Bigg) \\
&\leq \sum_{i\in[N]} \sum_{a_j^m\in\mathcal{K}_m} \sum_{m\in[M]} \mathbb{P}\Bigg(\left\| \mu_{i,j}^m - \hat{\mu}_{i,j,t}^{m,LS} \right\|_1 \geq \sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}} \Bigg)
\end{aligned}
$$

where the third equality is by De-Morgan's Law of sets. In the last inequality, we use the union bound to control the probability. Since each $\hat{\mu}_{i,j}^{m,LS} - \mu_{i,j}^m$ is a mean zero and $\frac{1}{2n_{i,j}^m}$-sub-Gaussian random variable, based on Lemma D.1, have $\mathbb{P}(\left\| \mu_{i,j}^m - \hat{\mu}_{i,j,t}^{m,LS} \right\|_1 \geq \sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}}) \leq \delta$. The overall bad event's probability's upper bound is

$$
\mathbb{P}(\overline{E}_t) \leq NK\delta \qquad \text{(C.14)}
$$

Based on our confidence width is less than 1, so $C = 1, \forall i \in [N]$. The expected regret from this bad event is not in the confidence interval at most

$$
NK\delta \cdot CT \leq NK \frac{1}{\sum_{i=1}^N Q_i T} T = \frac{NK}{Q} \qquad \text{(C.15)}
$$

This part's regret is negligible compared with the regret from Lemma F.2. In particular, if there is only one type and each firm has only one position to be filled. Thus, $Q = N$, the bad event's upper bounded probability will shrink to K, the number of workers to be explored. □

In this part, we provide the proof of MMTS's Bayesian regret upper bound.

F.4 Proof of Theorem 4.2

Theorem F.1. *When all firms follow the MMTS algorithm, the platform will incur the Bayesian total expected regret*

$$
\mathfrak{R}(T) \leq 8\log(QT)\sqrt{QK_{\max}}\sqrt{QT} + NK/Q \qquad \text{(C.16)}
$$

where $K_{\max} = \max\{K_1, ..., K_M\}, K = \sum_{m=1}^M K_m$.

Proof. We decompose the Bayesian total firm-optimal stable regret for all firms by

$$
\begin{aligned}
\mathfrak{R}(T) = \mathbb{E}_{\theta\in\Theta}\Bigg[\sum_{i=1}^N R_i(T,\theta) \Bigg] &= \mathbb{E}_{\theta\in\Theta}\Bigg[\sum_{i=1}^N \sum_{m=1}^M \sum_{t=1}^T \mu_{i,\overline{u}_i^m(t)}(t) - \sum_{i=1}^N \sum_{m=1}^M \sum_{t=1}^T \mu_{i,u_i^m}(t)|\theta \Bigg] \\
&= \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_{\theta\in\Theta}\Bigg[\sum_{m=1}^M (\mu_{i,\overline{u}_i^m(t)}(t) - \mu_{i,u_i^m}(t))|\theta \Bigg] \\
&= \mathbb{E}_{\theta\in\Theta}\Bigg[\sum_{t=1}^T \sum_{i=1}^N \sum_{m=1}^M \mathcal{I}_{i,t}^m|\theta \Bigg] \\
&= \mathbb{E}_{\theta\in\Theta}\Bigg[\sum_{t=1}^T \mathcal{I}_t|\theta \Bigg]
\end{aligned}
\qquad \text{(C.17)}
$$

where we define $\mathcal{I}_{i,t}^{m} = \mu_{i,\boldsymbol{\theta}}^{m}(\mathcal{A}_i^{m,*}) - \mu_{i,\boldsymbol{\theta}}^{m}(\mathcal{A}_{i,t}^{m})$ and $\mathcal{I}_t = \sum_{i=1}^{N}\sum_{m=1}^{M}\mathcal{I}_{i,t}^{m}$. Here $\mathcal{A}_i^{m,*}$ is the optimal matched workers for firm p_i of type m and $\mathcal{A}_{i,t}^{m}$ is the actual matched workers for firm p_i of type m at time t under the instance θ.

Based Lemma F.1, $\mathfrak{R}(T)$ is upper bounded by $\mathbb{E}\sum_{t=1}^{T}\left[C\mathbf{1}(\overline{E}_t)+\sum_{i=1}^{N}\sum_{m=1}^{M}\widetilde{W}_{i,\mathcal{F}_{i,t}^{m}}(\mathcal{A}_{i,t}^{m})\right]$. The first term, the sum of the bad event probability $\mathbb{E}\sum_{t=1}^{T}C\mathbf{1}(\overline{E}_t) = C\sum_{t=1}^{T}\mathbb{P}(\overline{E}_t)$, which is upper bounded by NK/Q based on Lemma F.4 and $C \leq 1$. The second term, the sum of confidence widths is upper bounded by $8Q\log(QT)\sqrt{TK_{\max}}$ based on Lemma F.2. Thus the Bayesian total regret is upper bounded by $8Q\log(QT)\sqrt{TK_{\max}} + NK/Q$. □

G INCENTIVE-COMPATIBILITY

In this section, we discuss the incentive-compatibility property of MMTS. That is, if one firm does not follow the MMTS when all other firms submit their MMTS preferences, that firm cannot benefit (matched with a better worker than his optimal stable matching worker) over a sublinear order. As we know, Dubins and Freedman (1981) discussed the *Machiavelli* firm could not benefit from incorrectly stating their true preference when there exists a unique stable matching. However, when one side's preferences are unknown and need to be learned through data, this result no longer holds. Thus, the maximum benefits that can be gained by the Machiavelli firm are under-explored in the setting of learning in matching. Liu et al. (2020) discussed the benefits that can be obtained by Machiavelli firm when other firms follow the centralized-UCB algorithm with the problem setting of one type of worker and quota equal one in the market.

We now show in CMCP, when all firms except one p_i submit their MMTS-based preferences to the matching platform, the firm p_i has an incentive also to submit preferences based on their sampling rankings in a *long horizon*, so long as the matching result do not have multiple stable solutions. Now we establish the following lemma, which is an upper bound of the expected number of pulls that a firm p_i can match with a m-type worker that is better than their optimal m-type workers, regardless of what preferences they submit to the platform.

Let's use $\mathcal{H}_{i,l}^{m}$ to define the achievable *sub-matching* set of $\mathbf{u}^m$ when all firms follow the MMTS, which represents firm p_i and $m-$ type worker a_l^m is matched such that $a_l^m \in \mathbf{u}_i^m$. Let $\Upsilon_{\mathbf{u}^m}(T)$ be the number of times sub-matching $\mathbf{u}^m$ is played by time t. We also provide the blocking triplet in a matching definition as follows.

Definition 3. *(Blocking triplet) A blocking triplet $(p_i, a_k, a_{k'})$ for a matching u is that there must exist a firm p_i and worker a_j that they both prefer to match with each other than their current match. That is, if $a_{k'} \in \mathbf{u}_i$, $\mu_{i,k'} < \mu_{i,k}$ and worker a_k is either unmatched or $\pi_{k,i} < \pi_{k,\mathbf{u}^{-1}(k)}$.*

The following lemma presents the upper bound of the number of matching times of p_i and a_l^m by time T, where a_l^m is a *super optimal m* $-$ type worker (preferred than all stable optimal *m* $-$ type workers under true preferences), when all firms follow the MMTS.

Lemma G.1. *Let $\Upsilon_{i,l}^{m}(T)$ be the number of times a firm p_i matched with a m-type worker such that the mean reward of a_l^m for firm p_i is greater than p_i's optimal match $\overline{\mathbf{u}}_i^m$, which is $\mu_{i,a_l^m}^{m} > \max_{a_j^m \in \overline{\mathbf{u}}_i^m}\mu_{i,j}^{m}$. Then the expected number of matches between p_i and a_l^m is upper bounded by*

$$\mathbb{E}[\Upsilon_{i,l}^{m}(T)] \leq \min_{S^m \in \mathcal{C}(\mathcal{H}_{i,l}^{m})} \sum_{(p_j, a_k^m, a_{k'}^m) \in S^m} \big(C_{i,j,k'}^{m}(T) + \frac{\log(T)}{d(\mu_{j,\overline{\mathbf{u}}_{i,\min}^{m}}, \mu_{j,k'})}\big),$$

where $\overline{\mathbf{u}}_{i,\min}^{m} = \operatorname*{argmin}_{a_k^m \in \overline{\mathbf{u}}_j^m} \mu_{i,k}^{m}$, and $C_{i,j,k'}^{m} = \mathcal{O}((\log(T))^{-1/3})$.

Then we provide the benefit (lower bound of the regret) of Machiavelli firm p_i can gain by not following the MMTS from matching with m-type workers. Let's define the *super worker reward gap* as $\overline{\Delta}_{i,l}^{m} = \max_{a_j^m \in \overline{\mathbf{u}}_i^m} \mu_{i,j}^{m} - \mu_{i,l}^{m}$, where $a_l^m \notin \overline{\mathbf{u}}_i^m$.

Theorem G.1. *Suppose all firms other than firm p_i submit preferences according to the MMTS to the centralized platform. Then the following upper bound on firm p_i's optimal regret for m-type workers* *holds:*

$$R_i^m(T,\theta) \geq \sum_{l:\overline{\Delta}_{i,l}^m<0} \overline{\Delta}_{i,l}^m \left[\min_{S^m \in \mathcal{C}(\mathcal{H}_{i,l}^m)} \sum_{(p_j, a_k^m, a_{k'}^m) \in S^m} \left(C_{i,j,k'}^m + \frac{\log(T)}{d(\mu_{j,\overline{\mathbf{u}}_{i,\min}^m}, \mu_{j,k'})} \right) \right] \tag{C.18}$$

where $\overline{\mathbf{u}}_{i,\min}^m = \operatorname*{argmin}_{a_k^m \in \overline{\mathbf{u}}_j^m} \mu_{i,k}^m$*, and* $C_{i,j,k'}^m = \mathcal{O}((\log(T))^{-1/3})$.

This result can be directly derived from Lemma G.1. Theorem G.1 demonstrates that there is no sequence of preferences that a firm can submit to the centralized platform that would result in negative optimal regret greater than $\mathcal{O}(\log T)$ in magnitude within type m. When considering multiple types together for firm p_i, this magnitude remains $\mathcal{O}(\log T)$ in total. Theorem G.1 confirms that, when there is a unique stable matching in type m, firms cannot gain significant advantage in terms of firm-optimal stable regret by submitting preferences other than those generated by the MMTS algorithm. An example is provided in Section 5.1 to illustrate this incentive compatibility property. Figure 2(b) illustrates the total regret, with solid lines representing the aggregate regret over all types for each firm, and dashed lines representing the regret for each type. It is observed that the type 1 regret of firm 1 is negative, owing to the inaccuracies in the rankings submitted by both firm 1 and firm 2. A detailed analysis of this negative regret pattern is given in Section I.2.

G.1 Proof of Incentive Compatibility

Lemma G.2. *Let* $\Upsilon_{i,l}^m(T)$ *be the number of times a firm* p_i *matched with a* m*-type worker such that the mean reward of* a_l^m *for firm* p_i *is greater than* p_i*'s optimal match* $\overline{u}_i^m$*, which is* $\mu_{i,a_l^m}^m > \max_{a_j^m \in \overline{u}_i^m} \mu_{i,j}^m$*. Then*

$$\mathbb{E}[\Upsilon_{i,l}^m(T)] \leq \min_{S^m \in \mathcal{C}(\mathcal{H}_{i,l}^m)} \sum_{(p_j, a_k^m, a_{k'}^m) \in S^m} \left(C_{i,j,k'}^m(T) + \frac{\log(T)}{d(\mu_{j,\overline{u}_{i,\min}^m}, \mu_{j,k'})} \right) \tag{C.19}$$

where $\overline{u}_{i,\min}^m = \operatorname*{argmin}_{a_k^m \in \overline{u}_j^m} \mu_{i,k}^m$*,* $C_{i,j,k'}^m = \mathcal{O}((\log(T))^{-1/3})$.

Proof. We claim that if firm p_i is matched with a *super optimal m* $-$ type worker a_l^m in any round, the matching u^m must be unstable according to true preferences from both sides. We then state that there must exist a m-type blocking triplet $(p_j, a_k^m, a_{k'}^m)$ where $p_j \neq p_i$.

We prove it by contradiction. Suppose all blocking triplets in matching u *only* involve firm p_i within $m -$ type worker. By Theorem 4.2 in (Abeledo and Rothblum, 1995), we can start from any matching u to a stable matching by iteratively satisfying blocking pairs in a *gender consistent* order, which means that we can provide a well-defined order to determine which blocking triplet should be satisfied (matched) first within preferences from firm p_i[6]. Doing so, firm p_i can never get a worse match than a_l^m since a blocking pair will let firm p_i match with a better $m -$ type worker than a_l^m, or become unmatched as the algorithm proceeds, so the matching will remain unstable. The matching will continue, which is a contradiction.

Hence there must exist a firm $p_j \neq p_i$ such that p_j is part of a blocking triplet in u when firm p_i is matched with $m -$ type worker a_l^m under the matching u. In particular, based on the Theorem 9 (Dubins-Freedman Theorem), firm p_j must submit its TS preference.

Let $L_{j,k,k'}^m(T)$ be the number of times firm p_j matched with $m -$ type worker $a_{k'}^m$ when the triplet $(p_j, a_k^m, a_{k'}^m)$ is blocking the matching provided by the centralized platform. Then by the definition

$$\sum_{u^m \in B_{j,k,k'}^m} \Upsilon_{u^m}(T) = L_{j,k,k'}^m(T) \tag{C.20}$$

By the definition of a blocking triplet, we know that if p_j is matched with $m -$ type worker $a_{k'}^m$ when the blocking triplet $(p_j, a_k^m, a_{k'}^m)$ is blocking, the TS sample must have a higher mean reward for $a_{k'}^m$

[6] This gender consistent requirement is to satisfy a blocking pair (p_j, a_k^m) and those blocking pairs can be ordered before we break their current matches if any, and then match p_j and a_k^m to get a new matching.

than a^m_k. In other words, we need to bound the expected number of times that the TS mean reward for $m-$ type worker $a^m_{k'}$ is greater than a^m_k. From (Komiyama et al., 2015), we know that the number of times that $(p_j, a^m_k, a^m_{k'})$ forms a blocking pair in Thompson sampling, is upper bounded by

$$\mathbb{E}L^m_{j,k,k'} \leq C^m_{i,j,k'}(T) + \frac{\log(T)}{d(\mu_{j,\overline{u}^m_{i,\min}}, \mu_{j,k'})} \tag{C.21}$$

where $\overline{u}^m_{i,\min} = \operatorname*{argmin}_{a^m_k \in \overline{u}^m_j} \mu^m_{i,k}$ and $C^m_{i,j,k'} = \mathcal{O}((\log(T))^{-1/3})$. The $d(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$ is the KL divergence between two Bernoulli distributions with expectation x and y.

The expected number of times $\Upsilon^m_{i,l}(T)$ a firm p_i matched with a $m-$ type worker such that the mean reward of a^m_l for firm p_i is greater than p_i's optimal match $\overline{u}^m_i$, which is equivalent to the expected number of times viat the achievable sub-matching set $\Upsilon_{u^m}(T)$ where $u^m \in \mathcal{H}^m_{i,l}$. So the result then follows from the identity

$$\mathbb{E}[\Upsilon^m_{i,l}(T)] = \sum_{u^m \in \mathcal{H}^m_{i,l}} \mathbb{E}\Upsilon_{u^m}(T) \tag{C.22}$$

Given a set $\mathcal{H}^m_{i,l}$ of matchings, we say a set S^m of triplets $(p_j, a^m_k, a^m_{k'})$ is a *cover* of $\mathcal{H}^m_{i,l}$ if

$$\bigcup_{(p_j, a^m_k, a^m_{k'}) \in S^m} B^m_{j,k,k'} \supseteq H^m_{i,l} \tag{C.23}$$

Let $\mathcal{C}(H^m_{i,l})$ denote the set of covers of $H^m_{i,l}$. Then

$$\begin{aligned}
\mathbb{E}[\Upsilon^m_{i,l}(T)] &= \mathbb{E}\sum_{u^m \in \mathcal{H}^m_{i,l}} \Upsilon_{u^m}(T) \\
&\leq \mathbb{E}\min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \Upsilon_{u^m}(T) \\
&= \min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \mathbb{E} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \Upsilon_{u^m}(T) \\
&= \min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \mathbb{E}L^m_{j,k,k'}(T) \\
&\leq \min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \left(C^m_{i,j,k'}(T) + \frac{\log(T)}{d(\mu_{j,k}, \mu_{j,k'})} \right) \\
&\leq \min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \left(C^m_{i,j,k'}(T) + \frac{\log(T)}{d(\mu_{j,\overline{u}^m_{i,\min}}, \mu_{j,k'})} \right)
\end{aligned} \tag{C.24}$$

where the first inequality is from the property of cover and we select the minimum cover S^m from $\mathcal{C}(\mathcal{H}^m_{i,l})$. And summation in the third line is equivalent to $\sum_{u^m \in B^m_{j,k,k'}}$. Based on Eq. (C.20), the third equality is obvious. From (Komiyama et al., 2015), we know the expected number of times of matching with the sub-optimal $m-$ type worker is upper bounded by Eq. (C.21). □

H Firm DA Algorithm with type and without type consideration

In this section, we present the DA algorithm with type consideration and without type consideration.

Algorithm 4: Firm DA Algorithm with type.

```
Input      : Type. firms set $\mathcal{N}$, workers set $\mathcal{K}_m, \forall m \in [M]$; firms to workers' preferences
             $\mathbf{r}_i^m, \forall i \in [N], \forall m \in [M]$, workers to firms' preferences $\boldsymbol{\pi}^m, \forall m \in [M]$; firms'
             type-specific quota $q_i^m, \forall i \in [N], \forall m \in [M]$, firms' total quota $Q_i, \forall i \in [N]$.
Initialize : Empty set $\mathcal{S} = \{\}$, empty sets $S^m = \emptyset, \forall m \in [M]$.
for $m = 1, ..., M$ do
    while $\exists$ A firm $p$ who is not fully filled with the quota $q^m$ and has not contacted every
      $m -$ type worker do
        Let $a$ be the highest-ranking worker in firm $p$'s preference, to whom firm $p$ has not yet
          contacted.
        Now firm $p$ contacts the worker $a$.
        if Worker $a$ is free then
            $(p, a)$ become matched (add $(p, a)$ to $S^m$).
        else
            Worker $a$ is matched to firm $p'$ (add $(p', a)$ to $S^m$).
            if Worker $a$ prefers firm $p'$ to firm $p$ then
                firm $p$ filled number minus 1 (remove $(p, a)$ from $S^m$).
            else
                Worker $a$ prefers firm $p$ to firm $p'$.
                firm $p'$ filled number minus 1 (remove $(p', a)$ from $S^m$).
                $(p, a)$ are paired (add $(p, a)$ to $S^m$).
    Update: Add $S^m$ to $\mathcal{S}$.
Output     : Matching result $\mathcal{S}$.
```

Algorithm 5: Firm DA Algorithm without type (Gale and Shapley, 1962).

```
Input      : Worker Types, firms set $\mathcal{N}$, workers set $\mathcal{K}_m, \forall m \in [M]$; firms to workers'
             preferences $\mathbf{r}_i^m, \forall i \in [N], \forall m \in [M]$, workers to firms' preferences $\boldsymbol{\pi}^m, \forall m \in [M]$;
             firms' type-specific quota $q_i^m, \forall i \in [N], \forall m \in [M]$, firms' total quota $Q_i, \forall i \in [N]$.
Initialize : Empty set $S$.
while $\exists$ A firm $p$ who is not fully filled with the quota $\tilde{Q}$ and has not contacted every worker do
    Let $a$ be the highest-ranking worker in firm $p$'s preference over all types of workers, to whom
      firm $p$ has not yet contacted.
    Now firm $p$ contacts the worker $a$.
    if Worker $a$ is free then
        $(p, a)$ become matched (add $(p, a)$ to $S$).
    else
        Worker $a$ is matched to firm $p'$ (add $(p', a)$ to $S$).
        if Worker $a$ prefers firm $p'$ to firm $p$ then
            firm $p$ filled number minus 1 (remove $(p, a)$ from $S$).
        else
            Worker $a$ prefers firm $p$ to firm $p'$.
            firm $p'$ filled number minus 1 (remove $(p', a)$ from $S$).
            $(p, a)$ are paired (add $(p, a)$ to $S$).
Output     : Matching result $S$.
```

I Experimental Details

In this section, we provide more details about the analysis of the negative regret, parameters, and large market.

I.1 Negative Regret Phenomenon

The occurrence of negative regret in multi-agent matching schemes presents an interesting phenomenon, contrasting the single-agent bandit problem wherein negative regret is non-existent.

Table 1: True matching rewards of two types of workers from two firms.

Mean ID	**Type**	**1**	**2**	**3**	**4**	**5**
$\boldsymbol{\mu}_1$	1	0.406	0.956	0.738	0.970	0.695
	2	0.932	0.241	0.040	0.657	0.289
$\boldsymbol{\mu}_2$	1	0.682	0.909	0.823	0.204	0.218
	2	0.303	0.849	0.131	0.886	0.428

Table 2: Estimated mean reward and variance of each type of worker in view of two firms. The bold font is to represent the firm's optimal stable matching. † represents the difference between the estimated mean and the true mean less than 1%. ‡ represents the difference is less than 1.5%.

Mean & Var	**Type**	**1**	**2**	**3**	**4**	**5**
$\hat{\boldsymbol{\mu}}_1$	1 (DS)	$0.533_{0.015}$	$\mathbf{0.943}^{\ddagger}_{0.000}$	$0.917_{0.035}$	$\mathbf{0.968}^{\dagger}_{0.000}$	$0.682^{\ddagger}_{0.003}$
	2 (SDE)	$\mathbf{0.950}_{0.000}$	$0.223_{0.000}$	$\mathbf{0.041}^{\dagger}_{0.000}$	$0.500_{0.208}$	$\mathbf{0.303}^{\ddagger}_{0.000}$
$\hat{\boldsymbol{\mu}}_2$	1 (DS)	$\mathbf{0.683}^{\dagger}_{0.000}$	$0.500_{0.035}$	$\mathbf{0.823}^{\dagger}_{0.000}$	$0.262_{0.037}$	$\mathbf{0.210}^{\dagger}_{0.000}$
	2 (SDE)	$0.083_{0.035}$	$\mathbf{0.851}^{\dagger}_{0.000}$	$0.124^{\dagger}_{0.001}$	$\mathbf{0.887}^{\dagger}_{0.000}$	$0.415^{\ddagger}_{0.001}$

In the context of the single-agent bandit problem, it is known that the best arm can be pulled, resulting in instantaneous regret that can attain zero but not take negative values. Conversely, in the multi-agent competing bandit problem, the oracle firm-optimal arm is determined by the true expected reward/utility, assuming knowledge of the true parameter μ^*. However, due to the imprecise estimation of rankings/parameters at each time step, an exact match with the oracle policy cannot be guaranteed. This discrepancy leads to varied outcomes for firms in terms of benefits (negative instantaneous regret) or losses (positive instantaneous regret) from the matching process. Instances arise where firms may strategically submit inaccurate rankings to exploit these matches, a phenomenon termed machiavelli/strategic behaviors. Nevertheless, over the long term, such strategic actions do not yield utility gains in accordance with our policy.

Furthermore, it is crucial to note that our matching solution remains a stable matching at each time step. This means that the stable matching remains independent of the negative regret generated by our policy, as stable matching is a short-term discrete metric, while regret serves as a long-term evaluation continuous metric.

I.2 Learning

In this section, we present the learning parameters of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of Example 1. Besides, we analyze which kind of pattern causes the non-optimal stable matching of Examples 1 and 2.

Findings from Example 1.

We show the posterior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in Figure 3. The first and second row represents the posterior distributions of firm 1 and firm 2 over two types of workers after T rounds interaction. The first and second columns in Figure 3 represent two firms' posterior distributions over type I and type II workers.

We find that the posterior distributions of the workers that firms most frequently match with exhibit a relatively sharp shape, indicating that firms can easily construct uncertainty sets over these workers. However, in some instances, the distributions are relatively flat, indicating a lack of exploration. This can be attributed to two possible reasons: (1) the workers in question are not optimal stable matches for the firms, and are thus abandoned early on in the matching process, such as firm 1's DS 1 and DS 5, or (2) the workers are optimal, but are erroneously ranked by the firms and subsequently blocked, such as firm 2's SDE 3. To further illustrate this, we present the posterior mean and variance in Table 2. The optimal stable matches for each firm are represented in bold, and the variance of the distributions is denoted by small font. Additionally, we use the dagger symbol to indicate when the difference between the posterior mean reward and true matching reward is less than 1% and 1.5%.

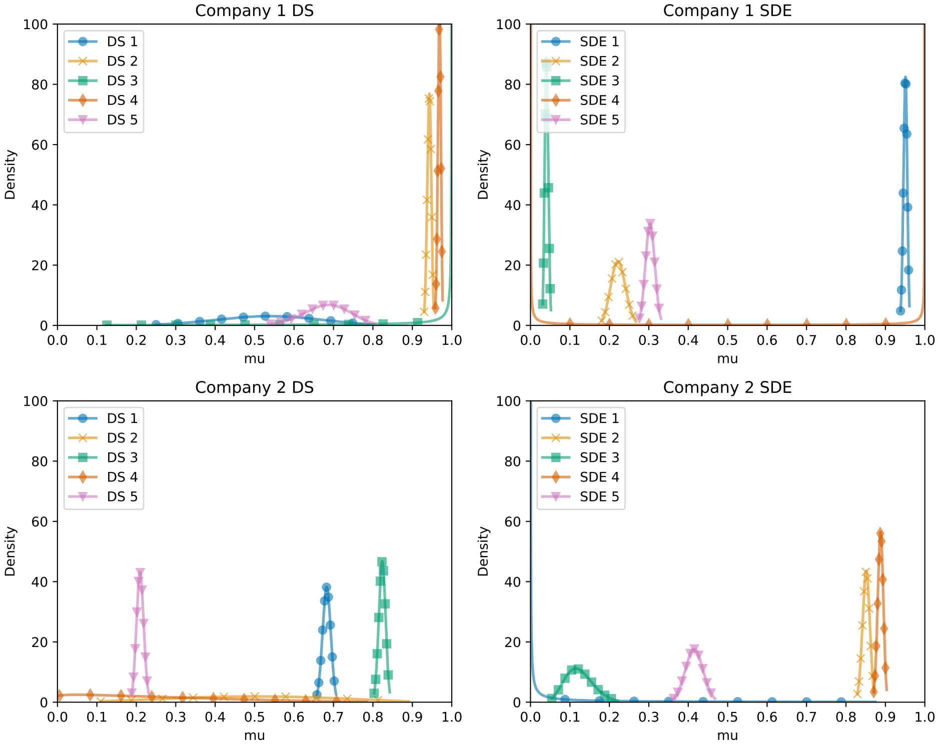

Figure 3: Posterior distribution of learning parameters for two firms in Example 1.

Pattern Analysis. We find that firm 1's type I matching in Figure 2(b), achieves a negative regret due to the high-frequency matching pattern of $\mathbf{u}_1 = \{[D_4, D_2, D_5], [S_1, S_5]\}$, and $\mathbf{u}_2 = \{[D_3, D_1], [S_4, S_2, S_3]\}$. That means firm 1 and firm 2 have a correct (stable) matching in the first match $\tilde{\mathbf{u}}_1 = \{[D_4, D_2], [S_1, S_5]\}, \tilde{\mathbf{u}}_2 = \{[D_3, D_1], [S_4, S_2]\}$. In the second match, they both need to compare worker D_5 and worker S_3, because all other workers are matched with firms or have been proposed in the first match. In Table 1, we find that two workers' true mean rewards for firm 1 are $\mu^1_{1,5} = 0.695, \mu^2_{1,3} = 0.040$ and two workers' estimated rewards for firm 1 are $\hat{\mu}^1_{1,5} = 0.682, \hat{\mu}^2_{1,3} = 0.041$. These two workers are pretty different and can be easily detected. So firm 1 has a high chance of ranking them correctly. However, two workers' true rewards for firm 2 are $\mu^1_{2,5} = 0.218, \mu^2_{2,3} = 0.131$, and two workers' estimated rewards for firm 2 are $\hat{\mu}^1_{1,5} = 0.210, \hat{\mu}^2_{1,3} = 0.124$. These workers are close to each other, where these two posteriors' distributions overlap a lot and can be checked in Figure 3. So firm 2 has a non-negligible probability to incorrectly rank S_3 ahead of D_5. Therefore, based on the true preference, firm 2 could match with S_3 and firm 1 matches with D_5 with a non-negligible probability rather than the optimal stable matching (p_1, S_3) and (p_2, D_5) by D_5 preferring firm 2.

The above pattern links to Section 3.2, incapable exploration, and Section G, incentive compatibility. Due to the insufficient exploration of S_3 and D_5, firm 2 may rank them incorrectly to get a match with S_3 rather than optimal D_3 and the regret gap is $\mu^1_{2,3} - \mu^2_{2,3} = 0.823 - 0.131 = 0.692$, which is a positive instantaneous regret. Due to the incorrect ranking from firm 2, firm 1 gets a final match with D_5 rather than optimal S_3, and suffers a regret gap $\mu^2_{1,3} - \mu^1_{1,5} = 0.040 - 0.695 = -0.655$, which is a negative instantaneous regret. Thus firm 1 benefits from firm 2's incorrect ranking and can achieve a total negative regret, as shown in Figure 2(b).

Findings from Example 2. In our analysis of the non-optimal stable matching in Example 2, we observed that both firms incurred positive total regret, shown in Figure 2(c). We find that the quota setting resulted in all workers of type II being assigned to firms in the first match. As a result, in the second match, the ranking submitted by firm 1 to the centralized platform did not affect firm 2's matching result for type II workers. This can be thought of as an analogy where firms are schools and workers are students. In the second stage of the admission process, school 2 would not participate in

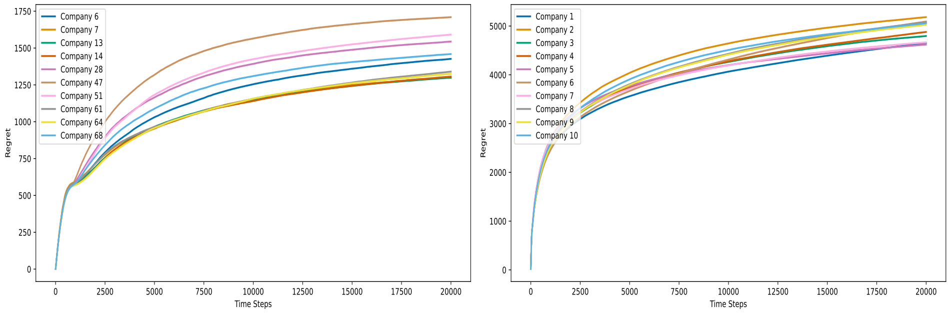

Figure 4: Left: 10 out of 100 randomly selected firms' total regret in Examples 3. Right: all firms' total regret in Example 4.

the competition for type II students, and its matching outcome would not be affected by the strategic behavior of other schools in the second stage, but rather by the strategic behavior of other schools in the first stage.

I.3 Large markets

In this part, we provide two large market examples to demonstrate the robustness of our algorithm. All preferences are randomly generated and all results are over 50 trials to take the average.

Example 3. We consider a large market composed of many firms ($N = 100$) and many workers ($K_1 = K_2 = 300$). Besides, we have $Q_1 = Q_2 = 3, q_1^1 = q_2^1 = q_2^1 = q_2^2 = 1$.

Example 4. We also consider a large market consisting of many workers, and each firm has a large, specified quota and an unspecified type quota. In this setting, $N = 10, M = 2, K_1 = K_2 = 500, Q_1 = Q_2 = 30, q_1^1 = q_2^1 = q_2^1 = q_2^2 = 10$.

Results. In Figure 4(a), we randomly select 10 out of 100 to present firms' total regret, and all those firms suffer sublinear regret. In Figure 4(b), we also show all 10 firms' total regret. Comparing Examples 3 and 4, we find that firms' regret in Example 3 is less than firms' regret from Example 4 because in Example 4, each firm has more quotas (30 versus 3), which demonstrates our findings from Theorem 4.2. In addition, we find there is a sudden exchange in Figure 4(a) nearby time $t = 1500$. We speculate this phenomenon is due to the small gap between different workers and the shifting of the explored workers.

J Additional Related Works

Multi-Agent Systems and Game theory. There are some papers considering the multi-agent in the sequential decision-making systems including the cooperative setting (Littman, 2001; González-Sánchez and Hernández-Lerma, 2013; Zhang et al., 2018; Perolat et al., 2018; Shi et al., 2022) and competing setting (Littman, 1994; Auer and Ortner, 2006; Zinkevich et al., 2007; Wei et al., 2017; Fiez et al., 2019; Jin et al., 2020). Zhong et al. (2021) study the multi-player general-sum Markov games with one of the players designated as the leader and the other players regarded as followers and establish the efficient RL algorithms to achieve the Stackelberg-Nash equilibrium.

Assortment Optimization. To maximize the number of matches between the two sides (customers and suppliers), the platform must balance the inherent tension between recommending customers more potential suppliers to match with and avoiding potential collisions. Ashlagi et al. (2022) introduce a stylized model to study the above trade-off. Motivated by online labor markets (Aouad and Saban, 2022) consider the online assortment optimization problem faced by a two-sided matching platform that hosts a set of suppliers waiting to match with a customer. Immorlica et al. (2021) consider a two-sided matching assortment optimization under the continuum model and achieve the optimized meeting rates and maximize the equilibrium social welfare. Rios et al. (2022) discuss the application of assortment optimization in dating markets. Shi (2022) studies the minimum communication needed for a two-sided marketplace to reach an approximately stable outcome with the transaction price.

Matching Markets. One strand of related literature is two-sided matching, which is a stream of papers that started in (Gale and Shapley, 1962). They propose the deferred acceptance (DA) algorithm (also known as the GS algorithm) with its application in the marriage problem and college admission problem. A series work (Knuth, 1976; Roth, 1982; Roth and Sotomayor, 1992; Roth, 2008) discuss the history of the DA algorithm and summarize theories about stability, optimality, and incentive compatibility, and finally provide its practical use and further open questions. In particular, Roth (1985); Sönmez (1997) propose that the college admissions problem is not equivalent to the marriage problem, especially when a college can manipulate its capacity and preference. Notably, in the hospital doctor matching example, since hospitals want diversity of specializations, or demographic diversity, or whatever, they care about the combination (group of doctors) they get. Roth (1986) state that when all preferences are strict, and hospitals (firms) have responsive preferences, the set of doctors (workers) employed and positions filled is the same at every stable match. However, when there exist *couples* in the preference list (not *responsive preference* (Klaus and Klijn, 2005)), which might make the set of stable matchings empty. Even when stable matchings exist, there need not be an optimal stable matching for either side. Later, Ashlagi et al. (2011) revisit this couple matching problem and provide the *sorted deferred acceptance algorithm* that can find a stable matching with high probability in large random markets. Biró et al. (2014) provide an integer programming model for hospital/resident problems with couples (HRC) and ties (HRCT). Manlove et al. (2017) release the HRC with minimal blocking pairs and show that if the preference list of every single resident and hospital is of length at most 2, their method can find a polynomial-time algorithm. Nguyen and Vohra (2018; 2022) find the stable matching in the nearby NRC problem, which is that the quota constraints are soft. Azevedo and Hatfield (2018); Che et al. (2019); Greinecker and Kah (2021) discuss the existence and uniqueness of stable matching with complementaries and its relationship with substitutable preferences in large economies. Besides, there are also papers considering stability and optimality of the refugee allocation matching (Aziz et al., 2018; Hadad and Teytelboym, 2022). Tomoeda (2018); Boehmer and Heeger (2022) consider a case that firms have hard constraints both on the minimum and maximum type-specific quotas and other type-specific quota consideration works.