Two-sided Competing Matching Recommendation Markets With Quota and Complementary Preferences Constraints

Yuantong Li [1] **Guang Cheng** [1] **Xiaowu Dai** [1 2]

Abstract

In this paper, we propose a new recommendation algorithm for addressing the problem of two-sided online matching markets with complementary preferences and quota constraints, where agents' preferences are unknown a priori and must be learned from data. The presence of mixed quota and complementary preferences constraints 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 new double matching technique to provide a stable matching outcome. Our theoretical analysis demonstrates the effectiveness of MMTS as it can achieve stability and has a total $\widetilde{\mathcal{O}}(Q\sqrt{K_{\max}T})$-Bayesian regret with high probability, which exhibits linearity with respect to the total firm's quota Q, the square root of the maximum size of available type workers $\sqrt{K_{\max}}$ and time horizon T. In addition, simulation studies also demonstrate MMTS' effectiveness in various settings. We provide code used in our experiments `https://github.com/Likelyt/Double-Matching`.

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 & 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. We consider the setting of matching markets with recommender systems, where preferences are usually unknown in the recommendation process due to the large volume of participants. 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 an example. In our proposed model, there is a set of agents (e.g., firms), each with a 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 for each type and a maximum quota of arms from all types. 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 *competing matching under complementary preference recommendation problem* (CMCPR).

The main contributions can be summarized as follows. Our first result is the formulation of CMCPR into a bandit learning framework as described in Lattimore & Szepesvári (2020). Using this framework, we propose a new algorithm, the Multi-agent Multi-type Thompson Sampling (MMTS), to solve CMCPR. Our algorithm builds on the strengths of Thompson Sampling (TS) (Thompson, 1933; Agrawal & Goyal, 2012; Russo et al., 2018) in terms of exploration and

[1]Department of Statistics and Data Science, UCLA; [2]Department of Biostatistics, UCLA. Correspondence to: Yuantong Li <yuantongli@ucla.edu>, Xiaowu Dai <dai@stat.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

further enhances it by incorporating a new *double matching* technique to find a stable solution for CMCPR, shown in Section 4.2. Unlike the upper confidence bound (UCB) algorithm, the 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 introduced double matching technique uses two stages of matching to satisfy both the type quota and total quota requirements. These two stages' matching mainly consists of using the deferred-acceptance (DA) algorithm from Gale & Shapley (1962).

Secondly, we provide the theoretical analysis of the proposed MMTS algorithm. Our analysis shows that MMTS achieves stability and enjoys incentive compatibility (IC). The proof of stability is obtained through a two-stage design of the double matching technique, and the proof of incentive compatibility is obtained through the regret lower bound. To the best of our knowledge, MMTS is the first algorithm to achieve stability and incentive compatibility in the CMCPR.

Finally, our theoretical results indicate that MMTS achieves a Bayesian regret that scales $\widetilde{\mathcal{O}}(\sqrt{T})$ and is near linear in terms of the total quota of all firms (Q). Besides, we find that the Bayesian 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, which is important for the large market. This is a more challenging setting than that considered in previous works such as Liu et al. (2020) and Jagadeesan et al. (2021), which only considers a single type of worker and a quota of one for each firm. To address these challenges, we use the eluder dimension (Russo & Van Roy, 2013) to measure the uncertainty set widths and bound the instantaneous regret for each firm, and use the 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 regret upper bound of MMTS.

The rest of this paper is organized as follows. Section 2 introduces basic concepts of CMCPR. Section 3 presents the challenges of this problem. Section 4 provides MMTS algorithm, its comparison with UCB-family algorithms, and shows the incapable exploration of the UCB algorithm in CMCPR. Section 5 provides the stability, regret upper bound, and the incentive-compatibility of MMTS. Section 6 shows the application of MMTS in simulations, including the distribution of learning parameters, and demonstrates the robustness of MMTS in large markets. Finally, Section 7 discusses related works.

2. Problem

We now describe the problem formulation of the **C**ompeting **M**atching under **C**omplementary **P**references **R**ecommendation problem (CMCPR).

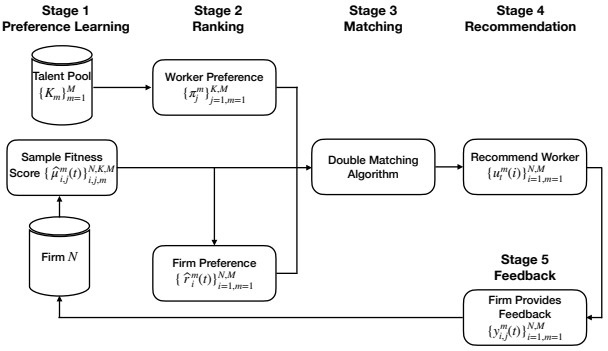

Figure 1. MMTS Algorithm for CMCPR with its application in the job market, including five stages: *preference learning*, *ranking construction*, *matching*, *recommendation*, *feedback collection*.

Notations. We define T as the time horizon and assume it is known in advance[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 vector.

2.1. Environment

The matching of workers and firms will be our running example throughout the paper. The organizer is the centralized platform, and the overall goal of the platform is to recommend the best-fit worker and match two-sided participants with their ideal objects over time. We first introduce seven elements in CMCPR.

(I) Participants. In the centralized platform, there are N firms (agents), denoted by $\mathcal{N} = \{p_1, p_2, ..., p_N\}$, and M types of workers (arms), represented $\mathcal{K}_m = \{a_1^m, a_2^m, ...a_{K_m}^m\}, m \in [M]$, where K_m is the number of m-th type workers and M is the total types.

(II) Quota. Agent p_i has a minimum quota q_i^m for m-type workers, and a maximum total quota Q_i (e.g., seasonal headcount in a company) and we assume $\sum_{i=1}^{M} q_i^m \leq Q_i$. Define the total market quota for all companies as $Q = \sum_{i=1}^{N} Q_i$ and the total number of available workers in the market as $K = \sum_{m=1}^{M} K_m$. We assume that $Q \ll K$ and T is relatively large.

(III) Two-sided Complementary Preferences. There are two kinds of preferences: workers to firms' preferences and firms to workers' preferences.

a. Preferences of m-type workers towards firms $\boldsymbol{\pi}^m$: $\mathcal{K}_m \mapsto \mathcal{N}$. We assume that there exist fixed preferences from workers to firms, and these preferences are known for the platform. For instance, workers submit their prefer-

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

ences for different firms on the platform. $\pi^m_{j,i}$ represents the rank for p_i from the view of a^m_j, and we assume that there are no ties in the rank orders, $\boldsymbol{\pi}^m_j \subseteq \{\pi^m_{j,1}, ..., \pi^m_{j,N}\}$. In other words, $\boldsymbol{\pi}^m_j$ is a subset of the permutation of $[N]$. And $\pi^m_{j,i} < \pi^m_{j,i'}$ implies that $m-$type worker a^m_j prefers firm p_i over firm $p_{i'}$ and as a shorthand, denoted as $p_i <^m_j p_{i'}$. This known worker-to-firm preference is a mild and common assumption in matching market literature (Liu et al., 2020; 2021; Li et al., 2022).

b. Preferences of firms towards $m-$type workers $\mathbf{r}^m : \mathcal{N} \mapsto \mathcal{K}_m$. The preferences of firms towards workers are fixed, but *unknown*. The goal of the platform is to infer these unknown preferences through historical matching data. We denote $r^m_{i,j}$ as the true rank of worker a^m_j in the preference list of firm p_i, and assume there are no ties. p_i's preferences towards workers is represented by $\mathbf{r}^m_i$, which is a subset of the permutation of $[\mathcal{K}_m]$. $r^m_{i,j} < r^m_{i,j'}$ implies that firm p_i prefers worker a^m_j over worker $a^m_{j'}$.

2.2. Policy

(IV) Matching Policy. $u^m_t(p_i) : \mathcal{N} \mapsto \mathcal{K}_m$ is a recommendation mapping function from p_i to $m-$ type workers at time t.

(V) Stable Matching and Optimal Matching. We introduce key concepts in matching fields (Roth, 2008).

Definition 1 (Blocking pair)**.** *A matching u is blocked by 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 $p_i >_j u^{-1}(a_j)$.*

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

Definition 3 (Valid Match)**.** *With true preferences from both sides, arm a_j is called a valid match of agent p_i if there exists a stable matching according to those rankings such that a_i and p_j are matched.*

Definition 4 (Agent Optimal Match)**.** *Arm a_j is an optimal match of agent p_i if it is the most preferred valid match.*

Given two-sided true preferences, the deferred-acceptance (DA) algorithm (Gale & Shapley, 1962) will provide a stable matching outcome. The matching result by the DA algorithm is always optimal for members of the proposing side, and we denote the agent-optimal policy as $\{\overline{u}^m_i\}^M_{m=1}$.

In CMCPR, it is worth mentioning that each firm has a minimum quota constraint $\mathbf{q}_i = [q^1_i, ..., q^M_i]$ for all type workers to fill and total quota cap is Q_i. Therefore, we define the concept of stability as the absence of "blocking pairs" across all types of workers and firms.[2]

(VI) Matching Score. If p_i is matched with a^m_j at time t, p_i provides a noisy reward $y^m_{i,j}(t)$ which is assumed to be the *true matching score* $\mu^m_{i,j}$ plus a noise term $\epsilon^m_{i,j}(t)$,

$$y^m_{i,j}(t) = \mu^m_{i,j} + \epsilon^m_{i,j}(t), \tag{1}$$

$\forall i, j, m, t \in [N], [K_m], [M], [T]$, where we assume that $\epsilon^m_{i,j}(t)$'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 $\mathbb{E}[\exp(\alpha\epsilon^m_{i,j}(t))] \leq \exp(\alpha^2\sigma^2/2)$.

(VII) Regret. Based on model (1), we denote the matching score for p_i as $\mathbf{y}^m_i(t) := \mathbf{y}_{i,u^m_t(p_i)}(t)$ in short. Define the *firm-optimal regret with m-type worker* for p_i as

$$R^m_i(T, \theta) := \sum_{t=1}^{T} [\mu_{i,\overline{u}^m_i} - \mu_{i,u^m_t(p_i)}(t) |\, \theta], \tag{2}$$

where denote θ as the sampled problem instance from the distribution Θ. $R^m_i(T, \theta)$ represents the total expected score difference between the policy $u^m_i := \{u^m_t(p_i)\}^T_{t=1}$ and the optimal policy $\overline{u}^m_i$ in hindsight.

As each firm have to recruit M types workers with total quota Q_i, the *total firm-optimal stable regret* for p_i is defined as

$$R_i(T, \theta) := \sum_{m=1}^{M} R^m_i(T, \theta). \tag{3}$$

Finally, define the *Bayesian social welfare gap* (BSWG) $\mathfrak{R}(T)$ as the expected regret over all firms and problem instance,

$$\mathfrak{R}(T) := \mathbb{E}_{\theta \in \Theta} \left[\sum_{i=1}^{N} R_i(T, \theta) \right]. \tag{4}$$

The goal of the centralized platform is to design a learning algorithm that achieves stable matchings through learning the firms' complementary preferences for multiple types of workers preciously from the previous matchings for a better recommendation. This is equivalent to designing an algorithm that minimizes BSWG $\mathfrak{R}(T)$.

3. 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 to solve 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

[2]The discussion of the feasibility of the stable matching in CMCPR is in Appendix A.

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 novel designed *double matching* (Algorithm 3) to marginalize complementary preferences and achieve stability. Our algorithm can efficiently learn a stable matching result using historical matching data, providing a practical solution to CMCPR.

Challenge 2: How to balance exploration and exploitation to achieve the sublinear regret? The platform must find a way to recommend the most suitable workers to firms to establish credibility among workers and firms to stay at the platform towards achieving optimal matching. Compared to traditional matching algorithms, the CMCPR is not a one-time recommendation algorithm but a *recycled* online recommendation matching algorithm with supply and demand consideration (workers and firms), which is more challenging as it requires more time to balance this trade-off. In addition, the classic UCB bandit methods could not function well in exploration and suffer sublinear regret demonstrated in Section 4.2. To overcome this challenge, we propose the use of a sampling algorithm, which allows for better exploration and achieves sublinear regret.

4. 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 minimize firms' Bayesian regret. We provide a description of MMTS and demonstrate the benefits of using the sampling method. The overall MMTS algorithm procedure is in Figure 1. The computational complexity of MMTS is in Appendix B.

4.1. Algorithm Description

The MMTS (Algorithm 1) is composed of five stages, *preference learning stage*, *ranking construction stage*, *double matching stage*, *collecting feedback stage*, and *updating belief stage*. At each matching step t, MMTS iterates these five steps.

Step 1: Preference Learning Stage. (Algorithm 2). For agent p_i, platform samples the mean feedback (reward) $\widehat{\mu}_{i,j}(t)$ of arm a_j^m from distribution $\mathcal{P}_j^m$ with estimated parameters $(\alpha_{i,j}^{m,t-1}, \beta_{i,j}^{m,t-1})$ from the historical matching data.

Step 2: Ranking Construction Stage. Then the platform sorts these workers within each type according $\{\widehat{\mu}_{i,j}(t)\}$ in descending order and gets the estimated rank $\widehat{\mathbf{r}}^m(t) = \{\widehat{\mathbf{r}}_i^m(t)\}_{i=1,m=1}^{N,M}$ where we denote $\widehat{\mathbf{r}}_i^m(t) = \{\widehat{\mathbf{r}}_{i,j}^m(t)\}_{j=1}^{K_m}$.

Step 3: Double Matching Stage. (Algorithm 3). With sampled mean reward $\widehat{\boldsymbol{\mu}}(t) := \{\widehat{\mu}_{i,j}^m(t)\}_{i=1,j=1,m=1}^{N,K_m,M}$, estimated ranks $\{\widehat{\mathbf{r}}^m(t)\}_{m=1}^M$, quota constraints $\{Q_i\}_{i=1}^N$, the double matching algorithm provides the final matching result with two-stage matchings.

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

Input : Time horizon T; firms' priors $(\boldsymbol{\alpha}_i^{m,0}, \boldsymbol{\beta}_i^{m,0}), \forall i, m \in [N], [M]$; workers' preference $\boldsymbol{\pi}^m, \forall m \in [M]$.

for $t \in \{1, ..., T\}$ **do**

STEP 1: PREFERENCE LEARNING STAGE
Sample estimated mean reward $\widehat{\boldsymbol{\mu}}_i^m(t)$ over all types of workers (Algo. 2)

STEP 2: RANKING CONSTRUCTION STAGE
Construct all firms' estimated rankings $\{\widehat{\mathbf{r}}_i^m(t)\}_{i=1,m=1}^{N,M}$ according $\widehat{\boldsymbol{\mu}}_i^m(t)$.

STEP 3: 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 4: RECOMMENDING AND COLLECTING FEEDBACK STAGE
Each firm receives its corresponding rewards from recommended all types of workers $\mathbf{y}_i^m(t)$.

STEP 5: UPDATING BELIEF STAGE
Based on received rewards, the platform updates firms' posterior belief.

end

The goal of the first match is to allow all firms to satisfy their minimum type-specific quota q_i^m first followed by sanitizing the status quo as a priori. The second match is to fill the left-over positions $\widetilde{Q}_i$ (defined below) for each firm and match firms and workers without type consideration.

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

b). Sanitize Quota: After the first match, the centralized platform sanitizes each firm's left-over quota $\widetilde{Q}_i = Q_i - \sum_{m=1}^M q_i^m$. If there exists a firm $p_i, s.t., \widetilde{Q}_i > 0$, then the platform will step into the second match. For those firms like p_i whose leftover quota is zero $\widetilde{Q}_i = 0$, they and their matched workers will skip the second match.

c). 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, the platform re-ranks the rest M

Algorithm 2 Preference Learning 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 $\widehat{\mu}_{i,j}^m(t) \sim \mathcal{P}(\alpha_{i,j}^{m,t-1}, \beta_{i,j}^{m,t-1}), \forall i, m, j \in [N], [M], [\mathcal{K}_m]$.

Sort: Sort estimated mean feedback $\widehat{\mu}_{i,j}^m(t)$ in descending order and get the estimated rank $\widehat{\mathbf{r}}_i^m(t)$.

Output: The estimated rank $\widehat{\mathbf{r}}_i^m(t)$ and the estimated mean feedback $\widehat{\boldsymbol{\mu}}_i^m(t), \forall i, m \in [N], [M]$.

Algorithm 3 Double Matching

Input : Estimated rank $\widehat{\mathbf{r}}(t)$, estimated mean $\widehat{\boldsymbol{\mu}}_i^m(t)$, 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

Given estimated ranks $\widehat{\mathbf{r}}(t)$ and all workers' preferences $\boldsymbol{\pi}^m$, the platform operates the firm-propose DA Algo and return the matching $\{\widetilde{u}_t^m(p_i)\}_{i=1,m}^{N,M}$.

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 $\widetilde{Q}_i \leftarrow Q_i - \sum_{m=1}^M q_i^m$ for firm p_i.

STEP 3: SECOND MATCH

if $\widetilde{\mathbf{Q}} \neq 0$ **then**

  Given left quota $\{\widetilde{Q}_i\}_{i\in[N]}$, estimated means $\widehat{\boldsymbol{\mu}}(t)$, and workers' preferences $\{\boldsymbol{\pi}^m\}_{m\in[M]}$, the platform runs the firm-propose DA and return the matching $\breve{u}_t(p_i)$.

else

  Set the matching $\breve{u}_t(p_i) = \emptyset$.

Output : The matching $u_t^m(p_i) \leftarrow \text{Merge}(\widetilde{u}_t^m(p_i), \breve{u}_t(p_i))$ for all firms.

types of workers who do not have a match in the first match for firms, 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 already or matched in Step 1. Then firm p_i gets the corresponding matched workers $\breve{u}_t(p_i)$ in the second match. Finally, the platform merges the first and second results to obtain a final matching $\mathbf{u}_t^m(p_i) = \text{Merge}(\widetilde{u}_t^m(p_i), \breve{u}_t(p_i)), \forall i, m \in [N], [M]$.

Step 4: Recommending and Collecting Feedback 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 5: Updating Belief Stage. After receiving these noisy rewards, the platform updates firms' belief (posterior) parameters as follows: $(\boldsymbol{\alpha}_i^{m,t}, \boldsymbol{\beta}_i^{m,t}) = \text{Update}(\boldsymbol{\alpha}_i^{m,t-1}, \boldsymbol{\beta}_i^{m,t-1}, \mathbf{y}_i^m(t)), \forall i \in [N], \forall m \in [M]$.

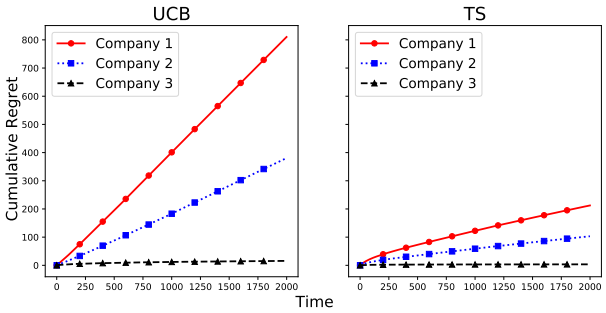

Figure 2. A comparison of centralized UCB and TS that demonstrates the incapable exploration of UCB.

4.2. Incapable Exploration

We show why the sampling method has an advantage over the UCB method in estimating worker ranks. We find that centralized UCB suffers linear firm-optimal stable regret in some cases and show it in Appendix C with detailed experimental setting and analysis.

Why sampling method 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 $\widehat{\mu}_j \approx \mu_j$, 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, 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 produces a linear regret for firm 1 and firm 2. However, the TS method achieves a sublinear regret in firm 1 and firm 2.

5. Properties of MMTS: Stability and Regret

Section 5.1 demonstrates the double matching algorithm can provide the stability property for CMCPR. Section 5.2 establishes the Bayesian regret upper bound for all firms when they follow the MMTS. Section 5.3 discusses the incentive-compatibility property of the MMTS.

5.1. Stability

In the following theorem, we show the double matching algorithm (Algo.3) provides a stable matching solution in the following theorem.

Theorem 5.1. *Given two sides' 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 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 & 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 not need to consider the type of workers designed by double matching. And the DA algorithm still provides 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. The detailed proof can be found in Appendix Section E. □

5.2. Bayesian Regret Upper Bound

Next, we provide MMTS's Bayesian total firm-optimal regret upper bound.

Theorem 5.2. *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 algorithm, MMTS, to effectively capture the interactions of multiple types of matching in CMCPR for the adaptation to the large market (K). 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 & 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.

5.3. Incentive-Compatibility

In this section, we discuss the incentive-compatibility property of MMTS. That is if one firm does not match the worker that MMTS (platform) recommended when all other firms follow MMTS recommended matching objects, which is equivalent to that firm submitting ranking preferences different from the sampled ranking list from MMTS, and we know that firm cannot benefit (matched with a better worker than his optimal stable matching worker) over a sublinear order. As we know, (Dubins & 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 firms 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 CMCPR, when all firms except one p_i accept their MMTS recommended workers from the matching platform, the firm p_i has an incentive also to follow the 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 workers they want to match.

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

Definition 5 (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* $a_{k'} \in u_i$*,* $\mu_{i,k'} < \mu_{i,k}$ *and worker* a_k *is either unmatched or* $\pi_{k,i} < \pi_{k,u^{-1}(k)}$*.*

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 5.1. *Let $\Upsilon^m_{i,l}(T)$ be the number of times 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 $\mu^m_{i,a^m_l} > \max_{a^m_j \in \overline{u}^m_i} \mu^m_{i,j}$. Then the expected number of matches between p_i and a^m_l is upper bounded by*

$$\mathbb{E}[\Upsilon^m_{i,l}(T)] \leq \min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \Big(C^m_{i,j,k'}(T) + \frac{\log(T)}{d(\mu_{j,\overline{u}^m_{i,\min}}, \mu_{j,k'})}\Big),$$

where $\overline{u}^m_{i,\min} = \operatorname{argmin}_{a^m_k \in \overline{u}^m_j} \mu^m_{i,k}$, *and* $C^m_{i,j,k'} = \mathcal{O}((\log(T))^{-1/3})$.

Then we show the benefit (lower bound of the regret) of Machiavelli firm p_i can gain by not following the MMTS recommended workers. Let's define the *super reward gap* as $\overline{\Delta}^m_{i,l} = \max_{a^m_j \in \overline{u}^m_i} \mu^m_{i,j} - \mu^m_{i,l}$, where $a^m_l \notin \overline{u}^m_i$.

Theorem 5.3. *Suppose all firms other than firm p_i follow the 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:*

$$R^m_i(T, \theta) \geq \sum_{l: \overline{\Delta}^m_{i,l} < 0} \overline{\Delta}^m_{i,l} \Bigg[\min_{S^m \in \mathcal{C}(\mathcal{H}^m_{i,l})} \sum_{(p_j, a^m_k, a^m_{k'}) \in S^m} \Bigg(C^m_{i,j,k'} + \frac{\log(T)}{d(\mu_{j,\overline{u}^m_{i,\min}}, \mu_{j,k'})} \Bigg) \Bigg].$$

This result can be directly derived from Lemma 5.1. Theorem 5.3 demonstrates that there is no sequence of preferences that a firm can manipulate and does not follow MMTS recommended workers that would achieve negative optimal regret and its absolute value greater than $\mathcal{O}(\log T)$. Considering M types together for firm p_i, this magnitude remains $\mathcal{O}(M \log T)$. Theorem 5.3 confirms that when there is a unique stable matching, firms cannot gain a significant advantage in terms of firm-optimal stable regret due to incorrect estimated preferences if others follow MMTS.

An example is provided in Section 6.1 to illustrate this incentive compatibility property. Figure 3(a) illustrates the total regret, with solid lines representing the aggregate regret over all types for each firm and dashed lines representing each type's regret. It is observed that the type I regret of p_1 is negative, owing to the inaccuracies in the rankings estimated for both p_1 and p_2. A detailed analysis of this negative regret pattern is given in Appendix Section I.1.

6. Experiments

In this section, we present simulation results to demonstrate the effectiveness of MMTS in learning firms' unknown preferences. The detailed experiment setup and the result can be found in Appendix Section I. Section 6.1 presents two examples to analyze the underlying causes of the novel phenomenon of negative regret (*gain benefit by matching with over-optimal workers*) and large market effect. Appendix Section I.1 showcases the distribution of learning parameters and provides insight into reasons for non-optimal stable matchings. Additionally, we demonstrate the robustness of MMTS in large markets in Appendix I.2. All simulation results are run in 100 trials.

6.1. Two Examples

Example 1. There are $N = 2$ firms, $M = 2$ types of workers, and there are $K_m = 5, \forall m \in [M]$. The quota q^m_i 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 workers to firms and from firms to workers are all randomly generated. Preferences from workers to firms' $\{\boldsymbol{\pi}^m\}_{m=1}^M$ are fixed and known. We use the data scientist (*D* or *DS*) and software developer engineer (*S* or *SDE*) as our example. The following are true preferences: $D_1 : p_1 \succ p_2, D_2 : p_1 \succ p_2, D_3 : p_2 \succ p_1, D_4 : p_1 \succ p_2, D_5 : p_2 \succ p_1, S_1 : p_1 \succ p_2, S_2 : p_1 \succ p_2, S_3 : p_2 \succ p_1, S_4 : p_2 \succ p_1, S_5 : p_1 \succ p_2$, and

$$\pi^1_1 : D_4 \succ D_2 \succ D_3 \succ D_5 \succ D_1,$$
$$\pi^2_1 : S_1 \succ S_4 \succ S_5 \succ S_2 \succ S_3,$$
$$\pi^1_2 : D_2 \succ D_3 \succ D_1 \succ D_5 \succ D_4,$$
$$\pi^2_2 : S_4 \succ S_2 \succ S_5 \succ S_1 \succ S_3.$$

The true matching scores of each worker for firms are sampled from $U([0, 1])$ and are available in Appendix Table 1. In addition, feedback $y^m_{i,j}(t)$ (0 or 1) provided by firms is generated by Bernoulli$(\mu^m_{i,j}(t))$. 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 $\alpha^{m,0}_{i,j} = \beta^{m,0}_{i,j} = 0.1, \forall i \in [N], \forall j \in [K_m], \forall m \in [M]$ to limit the strong impact of the prior belief. The update formula for each firm p_i at time t of the m-type worker a^m_j: $\alpha^{m,t+1}_{i,j} = \alpha^{m,t}_{i,j} + 1$ if the worker a^m_j is matched with the firm p_i, that is $a^m_j \in \mathbf{u}^m_t(p_i)$, and the provided score is $y^m_{i,j}(t) = 1$; otherwise $\alpha^{m,t+1}_{i,j} = \alpha^{m,t}_{i,j}$; $\beta^{m,t+1}_{i,j} = \beta^{m,t}_{i,j} + 1$ if the provided score is $y^m_{i,j}(t) = 0$, otherwise $\beta^{m,t+1}_{i,j} = \beta^{m,t}_{i,j}$. For other

unmatched pairs (firm, $m-$ type worker), parameters are retained.

Results. In Figure 3(a), we find that firms 1 and 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 estimated for firms, firm 1 benefits from this non-optimal matching result to achieve *negative* sublinear regret specifically for matching with type 1 workers often (blue dashed line).

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. 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, strategic actions do not yield utility gains in accordance with our policy.

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 still randomly generated. Therefore, the optimal 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 set in the same procedure as in Example 1.

Results. In Figure 3(b), 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 remaining quota would be all allocated to the type I workers 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

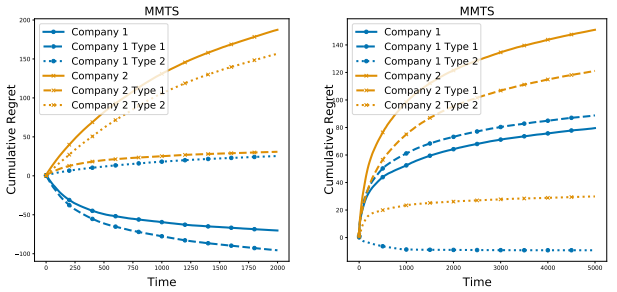

Figure 3. Firms and their sub-types regret for Example 1 and, firms and their sub-types regret for Example 2.

(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.

7. Related Works

This section reviews two-sided matching markets with unknown preferences, multi-agent systems, assortment optimization, and matching markets.

Two-sided Matching Market with Unknown Preferences. 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) applies it to the Markov matching market. Cen & Shah (2022) shows that if there is a transfer between agents, then the three desiderata (stability, low regret, and fairness) can be simultaneously achieved. Li et al. (2022) discusses the two-sided matching problem when the arm side has dynamic contextual information and preference is fixed from the arm side and proposes 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 & Jordan, 2021a;b; Dai et al., 2022; Kong et al., 2022; Kong & Li, 2023; Jagadeesan et al., 2022). In particular, Dai & Jordan (2021b) studies the college admission problem, provides an optimal strategy for agents, and shows its incentive-compatible property.

Multi-Agent Systems and Game theory. There are some papers considering the multi-agent in sequential decision-making systems including the cooperative setting (Littman, 2001; González-Sánchez & Hernández-Lerma, 2013; Zhang

et al., 2018; Perolat et al., 2018; Shi et al., 2022) and competing setting (Littman, 1994; Auer & Ortner, 2006; Zinkevich et al., 2007; Wei et al., 2017; Fiez et al., 2019; Jin et al., 2020). Zhong et al. (2021) studies the multi-player general-sum Markov games with one of the players designated as the leader and the other players regarded as followers and proposes 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) introduces a stylized model to study the above trade-off. Motivated by online labor markets Aouad & Saban (2022) considers 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) considers a two-sided matching assortment optimization under the continuum model to achieve the optimized meeting rates and maximize the equilibrium social welfare. Rios et al. (2022) discusses 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.

Two-sided Matching Markets with Known Preferences. One strand of related literature is two-sided matching, which is a stream of papers that started in Gale & Shapley (1962). They proposed the DA algorithm with its application in the marriage problem and college admission problem. A series of works (Knuth, 1976; Roth, 1982; Roth & Sotomayor, 1992; Roth, 2008) discuss the theories of the DA algorithm such as stability, optimality, and incentive compatibility, and provide the practical use. In particular, Roth (1985) and 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 and demographic diversity, they care about the combination (group of doctors) they get. Roth (1986) shows that if 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 exists *couples* in the preference list (not *responsive preference* (Klaus & Klijn, 2005)), it 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) revisits this couple matching problem and provides the *sorted deferred acceptance algorithm* that can find a stable matching with high probability in large random markets. Biró et al. (2014) provides an integer programming model for hospital/resident problems with couples (HRC) and ties (HRCT). Manlove et al. (2017) releases the HRC with minimal blocking pairs and shows 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 & Vohra (2018; 2022) find the stable matching in the nearby NRC problem, which is that the quota constraints are soft. Azevedo & Hatfield (2018); Che et al. (2019); Greinecker & 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 & Teytelboym, 2022). Tomoeda (2018); Boehmer & Heeger (2022) consider that firms have hard constraints both on the minimum and maximum type-specific quotas.

8. Conclusion and Future Work

In this paper, we proposed a new algorithm, MMTS to solve the CMCPR. MMTS builds on the strengths of TS for exploration and employs a double matching method to find a stable solution for complementary preferences and quota constraints. Through theoretical analysis, we show the effectiveness of the algorithm in achieving stability at every matching step under these constraints, achieving a $\tilde{\mathcal{O}}(Q\sqrt{K_{\max}T})$-Bayesian regret over time, and exhibiting the incentive compatibility 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 study 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 background information of agents, 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.

Acknowledgements

We would like to thank the area chair and anonymous referees for their constructive suggestions that improve the paper. Xiaowu Dai acknowledges support of CCPR as a part of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) population research infrastructure grant P2C-HD041022.

Impact Statement

This paper aims to advance the two-sided matching market by addressing complementarity and unknown preferences. Our work has potential societal implications, including promoting efficient matching for couples and enhancing diversity in the matching.

References

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

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

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

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

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

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

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

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

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

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

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

Boehmer, N. and Heeger, K. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kong, F. and Li, S. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1512–1522. SIAM, 2023.

Kong, F., Yin, J., and Li, S. Thompson sampling for bandit learning in matching markets. *arXiv preprint arXiv:2204.12048*, 2022.

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

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

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

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

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

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

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

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

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

Nguyen, T. and Vohra, R. Complementarities and externalities. *Online and Matching-Based Market Design*, 2022.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SUPPLEMENT TO TWO-SIDED COMPETING MATCHING RECOMMENDATION MARKETS WITH QUOTA AND COMPLEMENTARY PREFERENCES CONSTRAINTS

This supplement is organized as follows. In Section A, we discuss the feasibility of the matching problem with complementary preference and its corresponding assumption to secure a stable matching solution. In Section B, we show the computational complexity of MMTS. In Section C, we exhibit why the centralized UCB suffers insufficient exploration through a toy example. In Section D, we provide the fundamental Hoeffding concentration lemma for main theorems in this paper. 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. Finally, in Section I, we provide details of experiments and demonstrate the robustness of MMTS in large markets.

A. Feasibility of the Stable Matching

Assumption for the feasibility: For the two-sided finite market matching problem with complementary preferences, marginal preference is a sufficient condition for the feasibility. But for the large market, it requires more assumptions such as the substitutability and indifferences, etc,. The key difference between the finite and infinite market matching problem (Azevedo & Hatfield, 2018; Greinecker & Kah, 2021) lies in the agents availability. 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. In our case, in the finite market, 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 $a_1 > a_2$), 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 Section 7, if there exists couples in the preference list that cannot be marginalized, 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.

B. Complexity

Based on (Gale & 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 number of firm (N) and the number of workers $\sum_{m=1}^{M} K_m$.

C. Incapable Exploration

In this section, we show why the TS strategy has an advantage over the vanilla UCB 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]

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

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:

$$p_1 : a_1 \succ a_2 \succ a_3 \qquad a_1 : p_2 \succ p_3 \succ p_1$$
$$p_2 : a_2 \succ a_1 \succ a_3 \qquad a_2 : p_1 \succ p_2 \succ p_3$$
$$p_3 : a_3 \succ a_1 \succ a_2 \qquad a_3 : p_3 \succ p_1 \succ p_2$$

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 $\widehat{\mu}_3 \approx \mu_3$, 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, 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 4.2 Example - Insufficient Exploration

We set the true matching score 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 score. 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 4.2 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.

Proof. Let $\widehat{\mu}_{i,j,t}^{m,LS} = \frac{\sum_{s=1}^t \mathbf{1}(a_j^m \in \mathcal{A}_{i,s}^m) y_{i,j}^m(s)}{n_{i,j}^m(t)}$ 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:

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

Then 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.

[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.

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 & Shapley, 1962). The existence of stable matching has been given in (Gale & 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 $\widetilde{Q}_i$ 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.

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 roadmap of the regret decomposition and key steps to get Theorem 5.2. First, we define the history for firm p_i up to time t of type m as $H_{i,t}^m := \{\mathcal{A}_{i,1}^m, \mathbf{y}_{i,\mathcal{A}_{i,1}^m}^m(1), \mathcal{A}_{i,2}^m, \mathbf{y}_{i,\mathcal{A}_{i,2}^m}^m(2), ..., \mathcal{A}_{i,t-1}^m, \mathbf{y}_{i,\mathcal{A}_{i,t-1}^m}^m(t-1)\}$, composed by actions (matched workers) and rewards, where $\mathcal{A}_{i,t}^m := \mathbf{u}_t^m(p_i)$ is a set of workers (based on quota requirement q_i^m and Q_i) belong to m-type which is matched with firm p_i at time t, $\mathbf{y}_{i,\mathcal{A}_{i,t-1}^m}^m(t-1)$ are realized rewards when firm p_i matched with $m-$ type workers $\mathcal{A}_{i,t}^m$. Define $\widetilde{H}_{i,t} := \{H_{i,t}^1, H_{i,t}^2, ..., H_{i,t}^M\}$ as the aggregated interaction history between firm p_i and all types of workers up to time t.

[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.

Next, we define the *good event* for firm p_i when matching with $m-$ type worker at time t and the true mean matching score falls in the uncertainty set as $E_{i,t}^m = \{\boldsymbol{\mu}_{i,\mathcal{A}_{i,t}^m}^m \in \mathcal{F}_{i,t}^m\}$, where $\boldsymbol{\mu}_{i,\mathcal{A}_{i,t}^m}^m$ is the true mean reward vector of actually pulled arms (matched with $m-$ type workers) at time t for firm p_i, and $\mathcal{F}_{i,t}^m$ 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 $E_{i,t} = \bigcap_{m=1}^M E_{i,t}^m$, 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. 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,*

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

where $\widetilde{W}_{i,\widetilde{\mathcal{F}}_{i,t}^m}^m(\cdot) = \sum_{a_j^m \in \mathcal{A}_{i,t}^m} w_{i,\widetilde{\mathcal{F}}_{i,t}^m}^m(a_j^m)$ represents the sum of the element-wise value of uncertainty width at $m-$ type worker a_j^m. The uncertainty width $w_{i,\widetilde{\mathcal{F}}_{i,t}^m}^m(a_j^m) = \sup_{\bar{\mu}_i^m, \underline{\mu}_i^m \in \widetilde{\mathcal{F}}_{i,t}^m}(\bar{\mu}_i^m(a_j^m) - \underline{\mu}_i^m(a_j^m))$ is a worst-case measure of the uncertain about the mean reward of $m-$ type worker a_j^m. 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 $U_{i,t}^m(a) = \sup\{\mu_i^m(a) : \mu_i^m \in \mathcal{F}_{i,t}^m, a \in \mathcal{K}_m\}$ and $L_{i,t}^m(a) = \inf\{\mu_i^m(a) : \mu_i^m \in \mathcal{F}_{i,t}^m, a \in \mathcal{K}_m\}$, where μ_i^m is the mean reward function $\mu_i^m \in \mathcal{F}_{i,t}^m : \mathbb{R} \mapsto \mathbb{R}, \forall i \in [N], \forall m \in [M]$. Whenever $\mu_{i,\widetilde{\mathcal{A}}_i^m}^m \in \mathcal{F}_{i,t}^m$, the bounds $L_{i,t}^m(a) \leq \mu_{i,\widetilde{\mathcal{A}}_i^m}^m(a) \leq U_{i,t}^m(a)$ hold for all types of workers. Here we define $\mathcal{A}_{i,t}^m = \mathbf{u}_i^m(t)$ as the matched $m-$ type workers for firm p_i at time t and $\mathcal{A}_{i,t}^{m,*} = \overline{\mathbf{u}}_i^m(t)$ 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 $\mathcal{A}_i^{m,*} = \mathcal{A}_{i,t}^{m,*}$.

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* $\mathcal{F}_{i,j,t}^m := \{\mu_{i,j}^m \in \mathcal{F}_{i,j}^m : \left\|\mu_{i,j}^m - \widehat{\mu}_{i,j,t}^{m,LS}\right\|_1 \leq \sqrt{\beta_{i,j,t}^m}\}$, *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 & 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

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

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

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

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

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

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

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

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

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

if E_t holds. □

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

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

Here $\widehat{\mu}_{i,j,t}^{m,LS} = \frac{\sum_{s=1}^{t}\mathbf{1}(a_j^m \in \mathcal{A}_{i,s}^m)y_{i,j}^m(s)}{n_{i,j}^m(t)}$ is the estimated average reward for $m-$type worker a_j^m from the view point of firm p_i at time t, and $n_{i,j}^m(t)$ is the number of matched times up to time t of firm p_i with $m-$type worker a_j^m. Besides, we define $\widetilde{\beta}_{i,T} = \max_{a_j^m \in \mathcal{K}_m, m \in [M]}\beta_{i,j,T}^m$ 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 & 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_i^m.

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

$$\begin{aligned} \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{a_j^m \in \mathcal{A}_{i,t}^m}\mathbf{1}\Big(w_{i,\mathcal{F}_{i,t}^m}^m(a_j^m) > \epsilon\Big) &\leq \sum_{m=1}^{M}\sum_{a_j^m \in \mathcal{A}_{i,t}^m}\Big(\frac{4\beta_{i,j,T}^m}{\epsilon^2}+1\Big)\dim_E(\mathcal{F}_i^m, \epsilon) \\ &\leq \left(\frac{4\max_{a_j^m \in \mathcal{K}_m, m \in [M]}\beta_{i,j,T}^m}{\epsilon^2}+1\right)\sum_{m=1}^{M}|\mathcal{A}_{i,t}^m|\dim_E(\mathcal{F}_i^m, \epsilon), \end{aligned} \tag{C.11}$$

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

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

where $\widetilde{\beta}_{i,T} = \max_{a_j^m \in \mathcal{K}_m, m \in [M]}\beta_{i,j,T}^m$. Besides, we know that $Q_i = \sum_{m=1}^{M}|\mathcal{A}_{i,t}^m|$ 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 $\mathcal{F}_{i,j,t}^m := \{\mu_{i,j}^m \in \mathcal{F}_{i,j}^m : \left\|\mu_{i,j}^m - \widehat{\mu}_{i,j,t}^{m,LS}\right\|_1 \leq \sqrt{\beta_{i,j,t}^m}\}$ 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 score is not in the confidence interval we constructed is upper bounded by

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

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 $\widehat{\mu}_{i,j}^{m,LS} - \mu_{i,j}^m$ is a mean zero and $\frac{1}{2n_{i,j}^m}$-sub-Gaussian random variable, based on Lemma D.1, have $\mathbb{P}(\left\|\mu_{i,j}^m - \widehat{\mu}_{i,j,t}^{m,LS}\right\|_1 \geq \sqrt{\frac{\log(\frac{2}{\delta})}{n_{i,j}^m(t)}}) \leq \delta$. The overall bad event's probability's upper bound is

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

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

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

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 5.2

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

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

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

Proof. We decompose the Bayesian Social Welfare Gap for all firms by

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

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 $\mathbb{E}\sum_{t=1}^T\big[C\mathbf{1}(\overline{E}_t) + \sum_{i=1}^N\sum_{m=1}^M \widetilde{W}_{i,\mathcal{F}_{i,t}^m}(\mathcal{A}_{i,t}^m)\big]$. 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 $C\le 1$. 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 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 & 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 CMCPR, 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 6 (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*

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

where $\overline{\mathbf{u}}^m_{i,\min} = \underset{a^m_k \in \overline{\mathbf{u}}^m_j}{\operatorname{argmin}} \mu^m_{i,k}$, *and* $C^m_{i,j,k'} = \mathcal{O}((\log(T))^{-1/3})$.

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 $\overline{\Delta}^m_{i,l} = \max_{a^m_j \in \overline{\mathbf{u}}^m_i} \mu^m_{i,j} - \mu^m_{i,l}$, where $a^m_l \notin \overline{\mathbf{u}}^m_i$.

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:*

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

where $\overline{\mathbf{u}}^m_{i,\min} = \underset{a^m_k \in \overline{\mathbf{u}}^m_j}{\operatorname{argmin}} \mu^m_{i,k}$, *and* $C^m_{i,j,k'} = \mathcal{O}((\log(T))^{-1/3})$.

This result can be directly derived from Lemma 5.1. Theorem 5.3 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 5.3 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 6.1 to illustrate this incentive compatibility property. Figure 3(a) 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.1.

G.1. Proof of Incentive Compatibility

Lemma G.2. *Let $\Upsilon^m_{i,l}(T)$ be the number of times 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 $\mu^m_{i,a^m_l} > \max_{a^m_j \in \overline{u}^m_i} \mu^m_{i,j}$. Then*

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

where $\overline{u}^m_{i,\min} = \underset{a^m_k \in \overline{u}^m_j}{\operatorname{argmin}} \mu^m_{i,k}$, $C^m_{i,j,k'} = \mathcal{O}((\log(T))^{-1/3})$.

Proof. We claim that if firm p_i is matched with a *super optimal* $m-$ type worker a^m_l 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 $(p_j, a^m_k, a^m_{k'})$ 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 & 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^m_l since a blocking pair will let firm p_i match with a better $m-$ type worker than a^m_l, 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^m_l under the matching u. In particular, based on the Theorem 9 (Dubins-Freedman Theorem), firm p_j must submit its TS preference.

Let $L^m_{j,k,k'}(T)$ be the number of times firm p_j matched with $m-$ type worker $a^m_{k'}$ when the triplet $(p_j, a^m_k, a^m_{k'})$ is blocking

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

the matching provided by the centralized platform. Then by the definition

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

By the definition of a blocking triplet, we know that if p_j is matched with $m-$ type worker $a^m_{k'}$ when the blocking triplet $(p_j, a^m_k, a^m_{k'})$ is blocking, the TS sample must have a higher mean reward for $a^m_{k'}$ 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 $\overline{u}^m_{i,\min} = \underset{a^m_k \in \overline{u}^m_j}{\operatorname{argmin}} \mu^m_{i,k}$ 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-Proposing DA Algorithm with Type Consideration.

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).

  end

  Update: Add S^m to $\mathcal{S}$.

end

Output : Matching result $\mathcal{S}$.

Algorithm 5 Firm-Proposing DA Algorithm without Type Consideration (Gale & 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).

end

Output : Matching result S.

Table 1. True Matching Scores of two types of workers from two firms.

Mean ID	Type	1	2	3	4	5
μ_1	1	0.406	0.956	0.738	0.970	0.695
	2	0.932	0.241	0.040	0.657	0.289
μ_2	1	0.682	0.909	0.823	0.204	0.218
	2	0.303	0.849	0.131	0.886	0.428

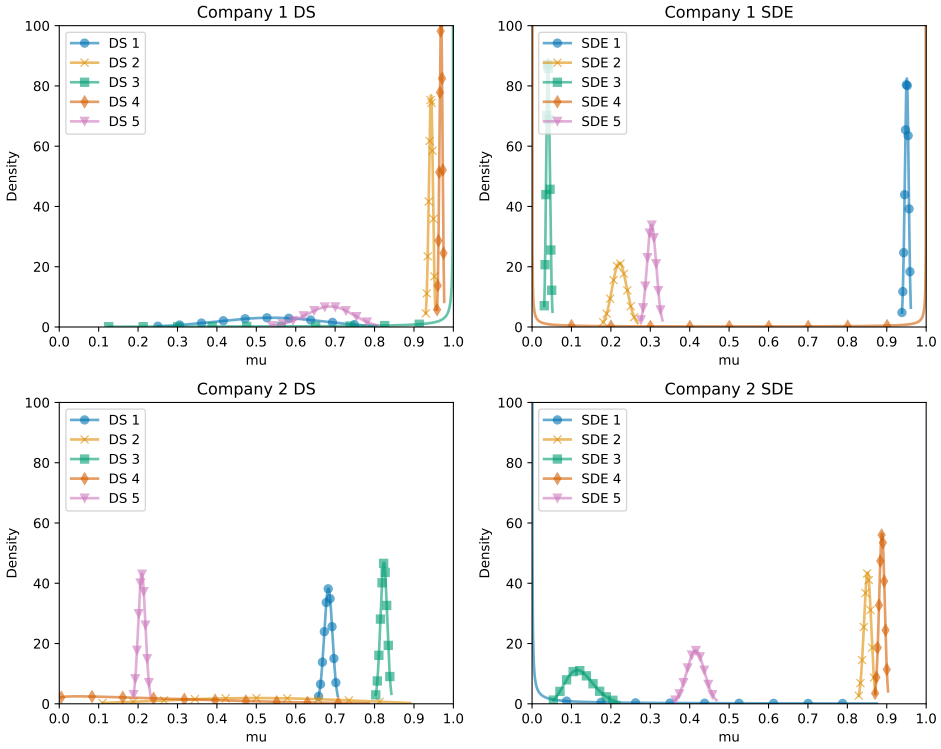

Figure 4. Posterior distribution of learning parameters for two firms in Example 1.

I. Experimental Details

In this section, we provide more details about the learned parameters and large market.

I.1. Learning Parameters

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.

We show the posterior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in Figure 4. 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 4 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

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)	$0.083_{0.035}$	$\mathbf{0.851}^{\dagger}_{0.000}$	$0.124^{\dagger}_{0.001}$	$\mathbf{0.887}^{\dagger}_{0.000}$	$0.415^{\ddagger}_{0.001}$

symbol to indicate when the difference between the posterior mean reward and true matching score is less than 1% and 1.5%.

Pattern Analysis. We find that firm 1's type I matching in Figure 3(a), 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 $\widehat{\mu}^1_{1,5} = 0.682, \widehat{\mu}^2_{1,3} = 0.041$. 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 $\widehat{\mu}^1_{1,5} = 0.210, \widehat{\mu}^2_{1,3} = 0.124$. These workers are close to each other, where these two posteriors' distributions overlap a lot and can be checked in Figure 4. 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 4.2, incapable exploration, and Section 5.3, 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 3(a).

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 3(b). 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 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.2. 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 $Q_1 = Q_2 = 3, q^1_1 = q^1_2 = q^1_2 = q^2_2 = 1$.

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, $N = 10, M = 2, K_1 = K_2 = 500, Q_1 = Q_2 = 30, q^1_1 = q^1_2 = q^1_2 = q^2_2 = 10$.

Results. In Figure 5(a), we randomly select 10 out of 100 to present firms' total regret, and all those firms suffer sublinear regret. In Figure 5(b), we also show all 10 firms' total regret. Comparing Examples 3 and 4, we find that firms' regret in

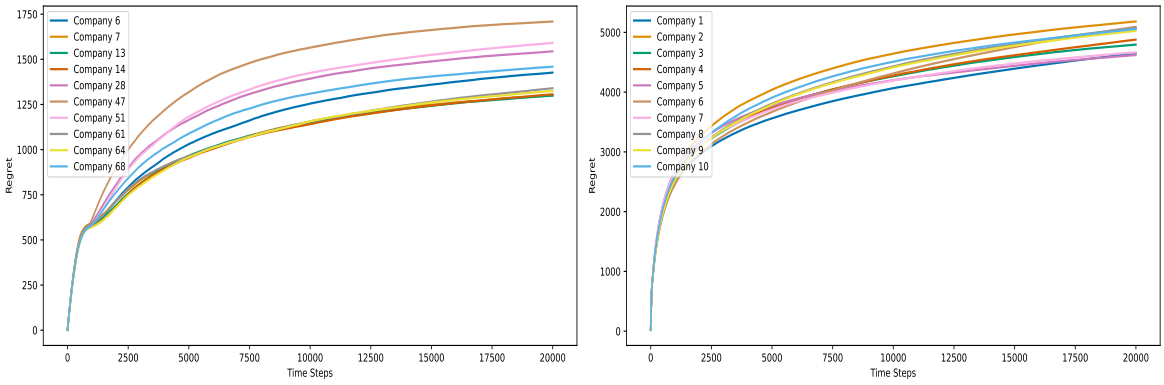

Figure 5. Left: 10 out of 100 randomly selected firms' total regret in Examples 3. Right: all firms' total regret in Example 4.

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 5.2. In addition, we find there is a sudden exchange in Figure 5(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.