Incentivized Exploration in Two-sided Matching Markets

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study *incentivized exploration* (IE) in centralized two-sided matching markets where all agents and arms are myopic human decision-subjects with preferences over their potential matches. The platform can leverage information asymmetry to encourage all sequentially arriving agents and arms to explore alternative options. In particular, we use inverse-gap weighting, a technique studied in reinforcement learning and contextual bandits, as the theoretical underpinning for our novel recommendation policy. We obtain the first set of results for incentivized exploration in two-sided matching markets with dual incentive-compatibility constraints and asymptotically match the regret guarantee for combinatorial semi-bandits.

1 Introduction

Consider an online job market where job applicants seek to get matched with employers in a one-to-one format, i.e., each job opening only accepts a single candidate. Each job applicant has their preference over which position they want to work in to utilize their skill set best. Similarly, employers want to match with candidates with well-documented track records who they can trust to perform well in the new job. This is a canonical example of the one-to-one matching problem studied by Gale and Shapley [1962]. While preference matching is ubiquitous, it may lead to self-imposed bias where job applicants only seek out employers they know beforehand, ignoring other options on the market. At the same time, employers also suffer from a lack of exploration as they are more favorable to prominent job applicants instead of expanding their search for the most suitable candidates. Moreover, in a large market, it is improbable that an employer can form an accurate preference ordering over job applicants without interacting with them first. Our goal is to *incentivize exploration* in a centralized matching market, where the platform provides recommendations for either the job applicants or the employees to explore alternative options. Such exploration is crucial to any learning algorithm that seeks to find the optimal matching in two-sided markets.

Overview of results. Our main contributions are as follows:

1. Prior work in incentivized exploration only considers the agents' incentives. Instead, this work considers the incentive-aware exploration problem in an online matching market from the perspectives of both agents and arms. See Appendix B for a detailed motivation.
2. We provide an end-to-end BIC algorithm with two components: 'warm-start' and accelerated exploration. Particularly, we develop a novel recommendation policy based on the inverse-gap weighting technique to accelerate exploration with near-optimal regret guarantees.
3. We provide numerical simulation on synthetic data and show that our end-to-end algorithm is both 1) incentive-compatible and 2) efficient in terms of regret minimization.

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

2 Preliminary

Notation. We write $[K] = \{1, 2, \cdots, K\}$ for $K \in \mathbb{N}^+$. We use subscripts i, j to denote different agents or arms, and superscript $t \in [T]$ to denote different time-steps.

We focus on an online two-sided matching market with time horizon T. At time-step $t \in [T]$, a fresh batch of N agents and N arms arrive and form N one-to-one matches. If they successfully match with some arms, the agents (and arms) report their shared utility to the platform and leave.

Reward formulation and Bayesian priors. We assume that the reward of each successful match is a bilinear function of the agent and the arm's profiles. Concretely, at time-step t, each agent of type i has their user profile $x_i^{(t)} \in \mathbb{R}^d$. Similarly, each arm of type j has profile vector $a_j^{(t)} \in \mathbb{R}^d$. Let $\Sigma \in \mathbb{R}^{d\times d}$ be a latent matrix with rank $r < d$. Then, the realized reward of a match (type i agent, type j arm) is:

$$r_{i,j}^{(t)} = r^{(t)}(x_i^{(t)}, a_j^{(t)}) \coloneqq (x_i^{(t)})^\top \Sigma a_j^{(t)} + \eta_{i,j}^{(t)} \tag{1}$$

where $\eta_{i,j}^{(t)} \sim \mathrm{subG}(\sigma)$. We write $\mu_{i.j} = x_i^\top \Sigma a_j$ to denote the expected reward of a match between agents of type i and arms of type j, and $\mu_{i,j}^{(0)}$ to denote the prior-mean reward. Wlog, we assume that $\forall i, j : \mu_{i,j} \in [0, 1]$. Henceforth, we write x_i and a_j to refer to agents of type i and arms of type j.

Preferences. We focus on the stylized setting with two types of agents and arms. Let $i, j \in [2]$ denote the type of agents and arms, respectively. We are interested in two sets of preferences: agent-to-arm and arm-to-agent. In our motivating example, job applicants want to be matched with compatible employers and employers prefer to be matched with applicants who can perform well. Wlog, we assume that the initial preference ordering is $\mu_{1,1}^{(0)} \ge \mu_{1,2}^{(0)} \ge \mu_{2,2}^{(0)}$ and $\mu_{1,1}^{(0)} \ge \mu_{2,1}^{(0)} \ge \mu_{2,2}^{(0)}$. That is, all agents prefer type 1 arms to type 2 arms, and all arms prefer type 1 agents to type 2 agents.

Incentive-compatibility. Absent incentives and coordination from the platform, the agents and arms match each other using their initial preferences. However, the platform wants to incentivize both the agents and the arms to explore different options to find the optimal matching and maximize the cumulative reward. In particular, at each time step t, the platform can broadcast a signal $\mathbf{s}^{(t)}$ as a recommendation to all agents and arms. By *direct revelation principle* [Myerson, 2018], this signal is equivalent to directly telling the agents which arm to match with, and vice versa.

Definition 2.1 (Two-sided Bayesian Incentive-Compatible Condition). *$\forall t \in [T]$, the platform's recommendation is $\epsilon-$*two-sided Bayesian Incentive-Compatible *(ϵ-BIC) for some $\epsilon > 0$ if it satisfies:*

$$\mathbb{E}[r_{i,j}^{(t)} | \mathrm{rec} = (x_i^{(t)}, a_j^{(t)})] - \sup_{\ell \in [N]} \mathbb{E}[r_{i,\ell}^{(t)} | \mathrm{rec} = (x_i^{(t)}, a_j^{(t)})] \ge \epsilon \tag{2}$$

$$\mathbb{E}[r_{i,j}^{(t)} | \mathrm{rec} = (x_i^{(t)}, a_j^{(t)})] - \sup_{\ell \in [N]} \mathbb{E}[r_{\ell,j}^{(t)} | \mathrm{rec} = (x_i^{(t)}, a_j^{(t)})] \ge \epsilon \tag{3}$$

Assumption 2.2 (Behavioral Assumption). *Agents and arms follow recommendations for any ϵ_0-BIC policy, for some fixed $\epsilon_0 > 0$. If one side rejects the recommendation, then both sides of the recommended (agent, arm) pair do have a match for that time-step and the platform receives a reward of* 0 *for that recommended pair. Both the agents and the arms are assumed to be* myopic*, i.e., they will choose the posterior best arms (agents) at the current time-step to match with.*

Reduction to combinatorial semi-bandits. Our first insight is to reduce the two-sided matching problem to a combinatorial semi-bandits problem. Consider the following mapping: at each time-step, the set of all feasible matches between agents and arms constitutes the action space $\mathcal{A} \subset \mathbb{R}^{N\times N}$. An *atom* $(x_i^{(t)}, a_j^{(t)})$ is a match between $x_i^{(t)}$ and $a_j^{(t)}$, and there are N^2 total atoms. An *action* $A^{(t)} \in \mathcal{A}$ at time-step t is the combination of matches at that round, where $\left\| A^{(t)} \right\|_1 \le N$. At each time-step t, a learner arrives at the platform, receives a recommendation for an action $A \in \mathcal{A}$, and chooses an action $A^{(t)} \in \mathcal{A}$. The platform and the learner both observe the reward of each atom in this arm (and nothing else). The algorithm's reward in this time-step is the total reward of these atoms.

Under this reduction, a few technical challenges differentiate our result from that of combinatorial semi-bandits. Particularly, it is unclear how to collect the 'warm-start' samples, which are input to any efficient incentivized exploration algorithm. For a detailed explanation, see Appendix B.

3 Incentivized exploration for two agents and two arms

In this section, we focus on the fundamental special case of incentivized exploration with two types of agents and arms to show the salient points of our analysis. In essence, the platform first incentivizes all agents and arms to match each other and collect samples from these matches. Then, the platform use these 'warm-start' samples to accelerate exploration and quickly converge to the optimal matching.

3.1 Initial exploration with Hidden Exploration

We present our first contribution, a BIC algorithm to collect the 'warm-start' samples, where the objective is to sample each atom, i.e., match between an agent and an arm, at least once and completes in T_0 time-steps for some T_0 determined by the prior. In the following algorithm, we show that in the 'worst case' with one 'explorable' atom initially, we can incentivize both the agents and the arms to explore different matches. Intuitively, given enough samples of the 'explorable' atom, we can split the remaining time-steps into phases such that in each phase, a new atom, i.e., a match between an agent and an arm that was previously not explorable, can be chosen by the learner upon receiving the principal's recommendation. The incentivized exploration technique within each phase builds on the approach from Mansour et al. [2015], which is defined for multi-armed bandits. However, the reward priors are highly correlated in two-sided matching markets, and the set of 'explorable' atoms can initially be of size 1. Furthermore, the intricate incentive interplay between agents and arms requires a more careful notion of which action to explore. Our technical contribution here is to provide a sequence of actions and prove that it is possible to incentivize both the agents and the arms to explore given some mild conditions on the posterior distribution of the reward for each atom.

We make the following non-degeneracy assumption: any action A_{cand} can be the posterior best action with a margin $\tau_{\mathcal{P}}$ and probability at least $\rho_{\mathcal{P}}$ after seeing at least $n_{\mathcal{P}}$ samples of the previous actions.

Assumption 3.1 (Fighting chance assumption)**.** *There exists number $n_{\mathcal{P}} \in \mathbb{N}$ and $\tau_{\mathcal{P}}, \rho_{\mathcal{P}} \in (0,1)$ determined by the prior $\mathcal{P}$ such that: for a sequence of actions $A^1_{\text{cand}}, \cdots, A^{N^2}_{\text{cand}}$ defined by $\text{NextCandidate}(\mathcal{A}, \mathcal{S}, \mathcal{P})$. Let $\mathcal{S}$ be the dataset containing exactly $k \in \mathbb{N}$ samples of each arm, then*

$$\Pr[X_i^k \geq \tau_{\mathcal{P}}] \geq \rho_{\mathcal{P}} \quad \forall i \in \mathcal{A} \text{ and } k \geq n_{\mathcal{P}}, \tag{4}$$

where $X_i^k = \min_{\text{arms } A \neq A_{\text{cand}}} \mathbb{E}[\mu_{A_{\text{cand}}} - \mu_A | \mathcal{S}]$

We state our initial sampling algorithm in Algorithm 1 and its theoretical guarantees in Theorem 3.2.

Algorithm 1: Initial sampling: Hidden Exploration

Input: Batch size $L \in \mathbb{N}$, target number of samples $k \in \mathbb{N}$, gap $C \in (0,1)$.

Initialize dataset $\mathcal{S} = \emptyset$;

The first k learners choose $A = \{(x_1, a_1)\}$ without recommendations. Let $\widehat{r}^k_{1,1}$ be the sample average of these rewards. Add these k samples to $\mathcal{S}$;

for each phase $\psi = 1$ to N^2 **do**

- $A^{(\psi)}_{\text{cand}} = \text{NextCandidate}(\mathcal{A}, \mathcal{S}, \mathcal{P})$;
- **if** $\widehat{r}^k_{1,1} \leq \mu^{(0)}_{A^{\psi}_{\text{cand}}} - C$ **then**
  - 'Exploit' action $A^* = A^{\psi}_{\text{cand}}$.
- **else**
  - 'Exploit' action $A^* = \{(x_1, a_1)\}$.
- From the set P of the next $L \cdot k$ learners, pick a set Q of k learners uniformly at random;
- Every learner $p \in P - Q$ is recommended the 'exploit' action A^*;
- Every learner $p \in Q$ is recommended action A_{cand}. Add the reward from all $p \in Q$ to $\mathcal{S}$.

Theorem 3.2. *Assuming Assumption 3.1 holds with constants $n_{\mathcal{P}}, \tau_{\mathcal{P}}, \rho_{\mathcal{P}}$. Then, Algorithm 1 is two-sided ϵ-BIC as long as the batch size L is at least*

$$L \geq 1 + \max\left\{\frac{2+2\epsilon}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}} - 2\epsilon}, \frac{2\epsilon}{\mu^{(0)}_{1,2} + \mu^{(0)}_{2,1} - \mu^{(0)}_{2,2} + \mathbb{E}[\Delta^k_{A^0, A_{2,2}} | \xi_3] \Pr[\xi_3] - 2\epsilon}\right\} \tag{5}$$

and completes in $T_0 = N^2 \cdot n_{\mathcal{P}} \cdot \frac{1+N^2}{\tau_{\mathcal{P}} \cdot \rho_{\mathcal{P}}}$ time-steps. All actions are sampled at least $n_{\mathcal{P}}$ times.

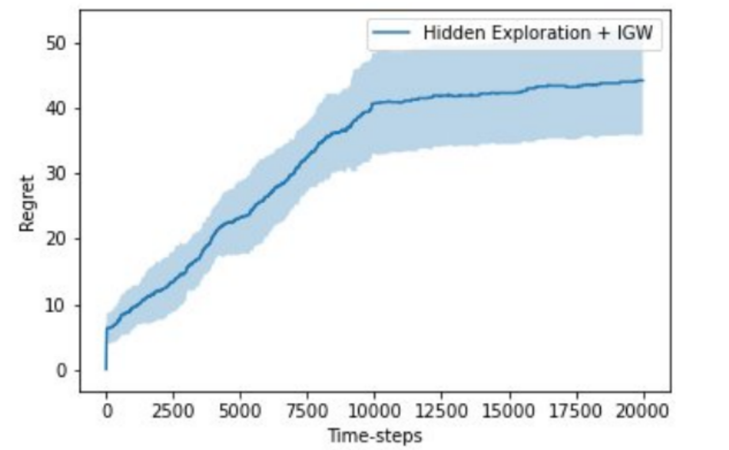


Figure 1: Regret using Algorithm 1 and Inverse Gap Weighting with time horizon $T = 20000$. Results are averaged over 10 runs, with the shaded region representing one standard error.

3.2 Accelerated Exploration with Inverse Gap Weighting

Given the data collected by Algorithm 1, the platform wants to accelerate exploration and converge to the optimal matching. The platform has to balance *exploitation*, i.e., recommending the empirical best match to minimize regret, and *exploration*, i.e., ensuring that the two-sided BIC condition holds. The theoretical underpinning of our recommendation policy at this stage is *inverse gap weighting*, i.e., recommending a match with probability inversely proportional to the reward gap between that match and the empirical best match. Formally, we let $b^{(t)} = \text{argmax}_{A \in \mathcal{A}} \hat{r}_A^{(t)}$ denote the empirical best action at time-step t. Then, the probability of an action A being recommended at time-step t is: $p_A^{(t)} = \begin{cases} \frac{1}{N^2 + \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})} & \text{if } A \neq b^{(t)} \\ 1 - \sum_{A \neq b^{(t)}} p_A^{(t)} & \text{otherwise} \end{cases}$, where the hyperparameter $\gamma > 0$ shows the tradeoff between exploration and exploitation. A smaller γ leads to more exploration, while a larger γ induces more exploitation. To ensure that γ is adaptive to the samples collected, we set $\gamma = C_0 \cdot N\sqrt{1/\phi^{(t)}}$, where $\phi^{(t)}$ is the mean squared error of the prediction at time-step t. Similar to Foster and Rakhlin [2020], we assume there exists an efficient regression-oracle that accurately compute $\phi^{(t)}$ at time-step t. With this recommendation policy, we state the theoretical guarantee for accelerated exploration:

Theorem 3.3 (Informal)**.** *Given sufficiently many 'warm-start' samples of all atoms, the inverse gap weighting recommendation policy is two-sided ϵ-BIC. The total regret during this stage is $O(N\sqrt{dT\log(T)})$, which asymptotically matches the optimal regret of combinatorial semi-bandits.*

4 Numerical Simulations

In this section, we complement our theoretical results with an experiment (Figure 1) to show incentive compatibility and regret minimization of our combined algorithm. For details, see Appendix D.

5 Conclusion and Future Work

In this work, we present the first results for incentivized exploration in two-sided matching markets, where the agents and arms are individuals with preferences over their matches. We characterize the incentive-compatibility constraints and provide a reduction to combinatorial semi-bandits. With this reduction, we present a BIC algorithm that collects 'warm-start' samples and accelerates exploration to minimize regret. In the future, we want to extend this work in several directions. First, we want to analyze the setting with more than two types of agents and arms. Moreover, we are working on experiments using synthetic and real-world datasets to support our theoretical findings.

References

D. Bergemann and S. Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, March 2019. doi: 10.1257/jel.20181489. URL `https://www.aeaweb.org/articles?id=10.1257/jel.20181489`.

B. Chen, P. Frazier, and D. Kempe. Incentivizing exploration by heterogeneous users. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 798–818. PMLR, 06–09 Jul 2018. URL `https://proceedings.mlr.press/v75/chen18a.html`.

J. Chen and K. Song. Two-sided matching in the loan market. *International Journal of Industrial Organization*, 31(2):145–152, 2013. ISSN 0167-7187. doi: https://doi.org/10.1016/j.ijindorg.2012.12.002. URL `https://www.sciencedirect.com/science/article/pii/S0167718712001245`.

X. Dai, Yuan, Qi, and M. I. Jordan. Incentive-aware recommender systems in two-sided markets, 2022.

D. J. Foster and A. Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles, 2020. URL `https://arxiv.org/abs/2002.04926`.

P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 5–22, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653. doi: 10.1145/2600057.2602897. URL `https://doi.org/10.1145/2600057.2602897`.

D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. ISSN 00029890, 19300972. URL `http://www.jstor.org/stable/2312726`.

X. Hu, D. D. Ngo, A. Slivkins, and Z. S. Wu. Incentivizing combinatorial bandit exploration, 2022.

N. Immorlica, J. Mao, A. Slivkins, and Z. S. Wu. Incentivizing exploration with unbiased histories. *CoRR*, abs/1811.06026, 2018. URL `http://arxiv.org/abs/1811.06026`.

N. Immorlica, J. Mao, A. Slivkins, and Z. S. Wu. Bayesian exploration with heterogeneous agents. In *The World Wide Web Conference*, WWW '19, page 751–761, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313649. URL `https://doi.org/10.1145/3308558.3313649`.

K.-S. Jun, R. Willett, S. Wright, and R. Nowak. Bilinear bandits with low-rank structure, 2019.

A. Kalvit, A. Slivkins, and Y. Gur. Incentivized exploration via filtered posterior sampling, 2024.

E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011. doi: 10.1257/aer.101.6.2590. URL `https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590`.

S. Kannan, M. Kearns, J. Morgenstern, M. Pai, A. Roth, R. Vohra, and Z. S. Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 369–386, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345279. doi: 10.1145/3033274.3085154. URL `https://doi.org/10.1145/3033274.3085154`.

M. Kasy and A. Teytelboym. Matching with semi-bandits. *The Econometrics Journal*, 26(1):45–66, 09 2022. ISSN 1368-4221. doi: 10.1093/ectj/utac021. URL `https://doi.org/10.1093/ectj/utac021`.

I. Kremer, Y. Mansour, and M. Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014. ISSN 00223808, 1537534X. URL `http://www.jstor.org/stable/10.1086/676597`.

Y. Li, G. Cheng, and X. Dai. Dynamic online recommendation for two-sided market with bayesian incentive compatibility, 2024a.

Y. Li, G. Cheng, and X. Dai. Two-sided competing matching recommendation markets with quota and complementary preferences constraints, 2024b.

Y. Mansour, A. Slivkins, and V. Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, EC '15, page 565–582, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334105. doi: 10.1145/2764468.2764508. URL https://doi.org/10.1145/2764468.2764508.

R. B. Myerson. *Revelation Principle*, pages 11646–11652. Palgrave Macmillan UK, London, 2018. ISBN 978-1-349-95189-5. doi: 10.1057/978-1-349-95189-5_2362. URL https://doi.org/10.1057/978-1-349-95189-5_2362.

Y. Papanastasiou, K. Bimpikis, and N. Savva. Crowdsourcing exploration. *Management Science*, 64 (4):1727–1746, 2018. doi: 10.1287/mnsc.2016.2697. URL https://doi.org/10.1287/mnsc.2016.2697.

J.-C. Rochet and J. Tirole. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1(4):990–1029, 06 2003. ISSN 1542-4766. doi: 10.1162/154247603322493212. URL https://doi.org/10.1162/154247603322493212.

M. Rysman and J. Wright. The economics of payment cards. *Review of Network Economics*, 13(3): 303–353, 2014. URL https://EconPapers.repec.org/RePEc:bpj:rneart:v:13:y:2014:i:3:p:303-353:n:4.

M. Sellke. Incentivizing exploration with linear contexts and combinatorial actions. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30570–30583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sellke23a.html.

A. Slivkins. Incentivizing exploration via information asymmetry. *XRDS*, 24(1):38–41, sep 2017. ISSN 1528-4972. doi: 10.1145/3123744. URL https://doi.org/10.1145/3123744.

A Related Work

Incentivized exploration The notion of incentivized exploration in this work has been introduced in [Kremer et al., 2014] and subsequently studied by [Mansour et al., 2015, Immorlica et al., 2018, 2019]. Recent work in incentivized exploration focuses on extending the framework beyond the classical multi-armed bandits setting. Notably, Hu et al. [2022], Sellke [2023] studied incentivized exploration using Thompson Sampling and allowing the Bayesian prior to be correlated across different arms. This framework is later generalized by [Kalvit et al., 2024] to allow for private agent types, informative recommendations, and correlated priors.

Incentivized exploration is related to the literature on information design [Kamenica and Gentzkow, 2011, Bergemann and Morris, 2019], where each time-step of incentivized exploration is essentially an instance of Bayesian persuasion, a central model in this literature. There exists a line of work orthogonal to ours that seeks to incentivize exploration via payment [Frazier et al., 2014, Kannan et al., 2017, Chen et al., 2018], time-discounted rewards [Papanastasiou et al., 2018]. For a detailed discussion, see Slivkins [2017]. Absent incentives, our model reduces to multi-armed bandits and its extension to bilinear bandits [Jun et al., 2019].

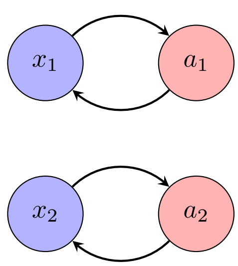
Two-sided matching market. The literature on two-sided matching market is first studied by the seminal work by [Gale and Shapley, 1962]. The two-sided market has many applications, ranging from streaming platforms to payment systems [Rysman and Wright, 2014] and loan market [Chen and Song, 2013]. For a broad overview of these applications, see Rochet and Tirole [2003]. The formulation of the two-sided matching problem as a combinatorial semi-bandits problem has been studied by Kasy and Teytelboym [2022]. There is a line of work on incentivized exploration in two-sided markets [Li et al., 2024b, Dai et al., 2022, Li et al., 2024a]. However, similar to other prior work in incentivized exploration, they only consider the agents' incentives in their algorithms. However, many real-world applications of two-sided matching markets have human decision-subjects on both sides whose incentives need to be taken into consideration when the platform designs a matching algorithm. In Appendix B, we describe a counterexample to illustrate the necessity of novel incentive mechanism designs for two-sided matching markets.

B Counterexample: One-sided incentive in matching market

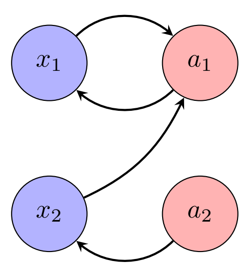
This section provides an example to show the need for dual BIC constraints in a two-sided matching market. Consider a stylized setting with two types of agents and arms and a time horizon of T. In the first T_0 time-steps, the platform runs a black-box recommendation algorithm such that, at the end of T_0 time-steps, the agents always follow the platform's recommendation and take the recommended arm. We show that there exists a problem instance where in the remaining $T - T_0$ time-steps, the algorithm incurs regret $\Omega(T - T_0)$.

We examine when a stable matching can happen without external incentives from the platform. The number of possible matchings between 2 agents and 2 arms is $2^4 = 16$ (each agent has 2 choices for which arm they prefer, and vice versa). Due to symmetry among the agents and the arms (e.g., a matching $\{(x_1, a_1), (x_2, a_2)\}$ is equivalent to the matching $\{(x_1, a_2), (x_2, a_2)\}$ by renaming the variables a_1 to a_2), there are 5 possible unique matchings between agents x_1, x_2 to arms a_1, a_2.

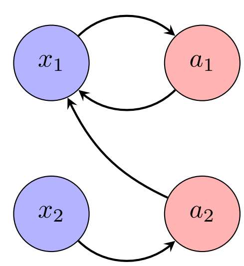
Among these unique matchings, only one is stable according to the initial preferences: Figure 2(a). If the optimal solution falls into this case (or its isomorphic forms), then the platform does not need to run an incentivized exploration algorithm to achieve optimal matching. However, for the remaining 4 possibilities, there always exists a possible realization of the rewards such that the initial preferences of either the agents or the arms will block an optimal matching (due to incompatible preference from either side) and any non-incentive-aware learning algorithm would incur linear regret.



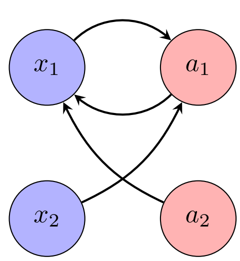
(a) Agent x_1 prefers arm a_1 (and vice versa), and agent x_2 prefers arm a_2 (and vice versa).



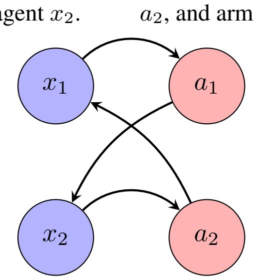
(b) Agent x_1 prefers arm a_1 (and vice versa). Agent x_2 prefers arm a_1, and arm a_2 prefers agent x_2.



(c) Agent x_1 prefers arm a_1 (and vice versa). Agent x_2 prefers arm a_2, and arm a_2 prefers agent x_1.



(d) Agent x_1 prefers arm a_1 (and vice versa). Agent x_2 prefers arm a_1, and arm a_2 prefers agent x_1.



(e) Agent x_1 prefers arm a_1, which prefers agent x_2. Agent x_2 prefers arm a_2, arm a_2 prefers agent x_1.

Figure 2: Possible unique matchings between 2 agents and 2 arms. Blue nodes on the left represent agents, and red nodes on the right represent arms. Arrow indicates that the start node prefers to be matched with the end node. Among all possible matchings, only the first case, where the matching forms two disjoint cyclic subgraphs, does not need the platform's interventions to have successful matches for all agents and arms. In any other cases, we can always find a blocking pair of nodes.

C Proofs of incentivized exploration for two types of agents and arms

C.1 Warm-start proofs

Proof of Theorem 3.2.

Proof. First, we show that agents of type 1 and arms of type 1 are willing to change their initial preference and follow the platform's recommendation. Then, we show that agents of type 2 and arms of type 2 will follow the recommendation and match each other.

Recommended action is $A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}$. For both agents of type 1 and arms of type 1 to change their initial preferences, we want to show that:

$$\mathbb{E}[\mu_{1,2} - \mu_{1,1} | A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}] \geq \epsilon \tag{6}$$

and

$$\mathbb{E}[\mu_{2,1} - \mu_{1,1} | A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}] \geq \epsilon \tag{7}$$

Combining these conditions, we instead will prove the following:

$$\mathbb{E}[\mu_{1,2} + \mu_{2,1} - 2\mu_{1,1)} | A^{(t)} = \{(x_1, a_2), (x_2, a_1)\}] \geq 2\epsilon \tag{8}$$

Let $A^0 = \{(x_1, a_1), (x_1, a_1)\}$ be a dummy action whose reward is twice the reward from choosing the prior-best atom (x_1, a_1). Then, our goal is to show that $\mathbb{E}[\mu_A - \mu_{A^0} | A^{(t)} = A] \geq \epsilon$.

Define the following two events:

$$\xi_1 = \{\text{exploit: } \mathbb{E}[\mu_A - \mu_{A^0} | S^k_{A^0}] > 0\} \tag{9}$$

and

$$\xi_2 = \{\text{explore: } \mathbb{E}[\mu_A - \mu_{A^0} | S^k_{A^0}] \leq 0 \text{ and selected for exploration}\} \tag{10}$$

Then, we can write

$$E[\mu_A - \mu_{A^0} | A^{(t)} = A] \geq \mathbb{E}[\mu_A - \mu_{A^0} | \xi_1] \Pr[\xi_1] + \mathbb{E}[\mu_A - \mu_{A^0} | \xi_2] \Pr[\xi_2]$$

Let $\Delta^k_{A,A^0} \coloneqq \mathbb{E}[\mu(A) - \mu(A^0) | S^k_{A^0}]$. Then, we have:

$$\begin{aligned}
\Pr[\xi_2] &= \Pr\left[\mathbb{E}[\mu_A - \mu_{A^0} | S^k_{A^0}] \leq 0 \text{ and selected for exploration}\right] \\
&= \Pr[\Delta^k_{A,A^0} \leq 0] \Pr[\text{selected} | \Delta^k_{A,A^0} \leq 0] \\
&= \frac{1}{L} \cdot \Pr[\Delta^k_{A,A^0} \leq 0]
\end{aligned}$$

where the first equality is by definition and the second equality is due to Δ^k_{A,A^0} being independent of the event that the learner is selected for exploration. Then, we can write

$$\begin{aligned}
&\mathbb{E}[\Delta^k_{A,A^0} | \xi_2] \Pr[\xi_2] \\
&= \mathbb{E}[\Delta^k_{A,A^0} | \Delta^k_{A,A^0} \leq 0 \text{ and selected}] \Pr[\Delta^k_{A,A^0} \leq 0] \cdot \frac{1}{L} \\
&= \mathbb{E}[\Delta^k_{A,A^0} | \Delta^k_{A,A^0} \leq 0] \Pr[\Delta^k_{A,A^0} \leq 0] \cdot \frac{1}{L}
\end{aligned}$$

Hence, the left-hand side of the dual BIC condition is

$$
\begin{aligned}
&\mathbb{E}[\Delta^k_{A,A^0}|A^{(t)}=A]\Pr[A^{(t)}=A]\\
&=\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\Pr[\Delta^k_{A,A^0}>0]\\
&+\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}\text{ and selected}]\Pr[\Delta^k_{A,A^0}<0\text{ and selected}]\\
&=\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\cdot Pr[\Delta^k_{A,A^0}>0]\\
&+\frac{1}{L}\cdot\mathbb{E}[\Delta^k_{A,A^0}\le 0|\Delta^k_{A,A^0}\le 0]\\
&=\left(1-\frac{1}{L}\right)\cdot\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\cdot\Pr[\Delta^k_{A,A^0}>0]\\
&+\frac{1}{L}\cdot\left(\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\cdot\Pr[\Delta^k_{A,A^0}>0]+\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}\le 0]\cdot\Pr[\Delta^k_{A,A^0}\le 0]\right)\\
&=\left(1-\frac{1}{L}\right)\cdot\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\cdot\Pr[\Delta^k_{A,A^0}>0]+\frac{1}{L}\cdot\mathbb{E}[\Delta^k_{A,A^0}]\\
&=\frac{L-1}{L}\cdot\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\cdot\Pr[\Delta^k_{A,A^0}>0]+\frac{1}{L}\cdot\left(\mu^0_A-\mu^0_{A^0}\right)
\end{aligned}
$$

For the dual BIC condition to hold, we can set

$$
\begin{aligned}
&\frac{L-1}{L}\cdot\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\Pr[\Delta^k_{A,A^0}>0]+\frac{1}{L}\cdot(\mu^{(0)}_A-\mu^{(0)}_{A^0})\ge 2\epsilon\\
&\iff L\ge\frac{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\Pr[\Delta^k_{A,A^0}>0]-(\mu^{(0)}_A-\mu^{(0)}_{A^0})}{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\Pr[\Delta^k_{A,A^0}>0]-2\epsilon}\\
&\iff L\ge 1-\frac{\mu^{(0)}_A-\mu^{(0)}_{A^0}-2\epsilon}{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\Pr[\Delta^k_{A,A^0}>0]-2\epsilon}\\
&\iff L\ge 1+\frac{\mu^{(0)}_{A^0}-\mu^{(0)}_A+2\epsilon}{\mathbb{E}[\Delta^k_{A,A^0}|\Delta^k_{A,A^0}>0]\Pr[\Delta^k_{A,A^0}>0]-2\epsilon}
\end{aligned}
$$

The expression above can be simplified by using definitions of $\tau_{\mathcal{P}},\rho_{\mathcal{P}}$ and observing that $\mu^{(0)}_{A^0}-\mu^{(0)}_A\ge 2$ to get

$$
L\ge 1+\frac{2+2\epsilon}{\tau_{\mathcal{P}}\cdot\rho_{\mathcal{P}}-2\epsilon}\tag{11}
$$

Recommended action is $A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}$. We want to show that all agents and arms will comply with this recommendation. That is, we want to show

$$
\begin{aligned}
\mathbb{E}[\mu_{2,2}-\mu_{2,1}|A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}]\ge\epsilon\\
\mathbb{E}[\mu_{2,2}-\mu_{1,2}|A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}]\ge\epsilon\\
\mathbb{E}[\mu_{1,1}-\mu_{1,2}|A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}]\ge\epsilon\\
\mathbb{E}[\mu_{1,1}-\mu_{2,1}|A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}]\ge\epsilon
\end{aligned}
$$

Let $A_{2,2}=\{(x_2,a_2),(x_2,a_2)\}$ and $A_{1,1}=\{(x_1,a_1),(x_1,a_1)\}$ be a pair of dummy actions with reward twice that of atom (x_2,a_2) and (x_1,a_1), respectively. Let $A^0=\{(x_1,a_2),(x_2,a_1)\}$ denote the prior-best actions for x_2 and a_2. Then, we can combine these conditions and show that:

$$
\begin{aligned}
\mathbb{E}[\mu_{A_{2,2}}-\mu_{A^0}|A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}]\ge 2\epsilon\\
\mathbb{E}[\mu_{A_{1,1}}-\mu_{A^0}|A^{(t)}=\{(x_1,a_1),(x_2,a_2)\}]\ge 2\epsilon
\end{aligned}
$$

First, we consider the incentives of x_1 and a_1. We have:

$$
\begin{aligned}
&\mathbb{E}[\mu_{A_{1,1}}-\mu_{A^0}|\neg\text{explore}]\Pr[\neg\text{explore}]\\
&\mathbb{E}[\mu_{A_{1,1}}-\mu_{A^0}]-\mathbb{E}[\mu_{A_{1,1}}-\mu_{A^0}|\text{explore}]\Pr[\text{explore}]\\
&=2\mu^{(0)}_{1,1}-(\mu^{(0)}_{1,2}+\mu^{(0)}_{2,1})+\mathbb{E}[\mu_{A^0}-\mu_{A_{1,1}}|\text{explore}]\Pr[\text{explore}]
\end{aligned}
$$

Since the first term is non-negative according to the initial preference ordering, it suffices to show that $\mathbb{E}[\mu_{A^0} - \mu_{A_{1,1}}|\text{explore}] \Pr[\text{explore}] \geq 2\epsilon$. This inequality holds from the previous analysis for recommending $A_{\text{cand}} = \{(x_1, a_2), (x_2, a_1)\}$.

Then, we consider the incentives of x_2 and a_2. By construction, when agent x_2 receives a recommendation for arm a_2, they can infer that they are not in the explore group. Hence, it suffices to show that $\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0}|\neg\text{explore}] \Pr[\neg\text{explore}] \geq 2\epsilon$. We have:

$$\begin{aligned}
&\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0}|\neg\text{explore}] \Pr[\neg\text{explore}] \\
&= \mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0}] - \mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0}|\text{explore}] \Pr[\text{explore}] \\
&= (2\mu_{2,2}^{(0)} - \mu_{2,1}^{(0)} - \mu_{1,2}^{(0)}) + \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|\text{explore}] \Pr[\text{explore}]
\end{aligned}$$

Define the following events:

$$\begin{aligned}
\xi_3 &= \{\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|S_{1,1}^k] > 0\} \\
\xi_4 &= \{\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|S_{1,1}^k] \leq 0\}
\end{aligned}$$

Then, we can write:

$$\begin{aligned}
&\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|\text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|\xi_3] \Pr[\xi_3] + \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|\xi_4] \Pr[\xi_4]
\end{aligned}$$

Let $\Delta_{A^0,A_{2,2}}^k = \mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|S_{1,1}^k]$. Then, we have:

$$\begin{aligned}
\Pr[\xi_3] &= \Pr[\Delta_{A^0,A_{2,2}}^k \leq 0|\text{selected for exploration}] \Pr[\text{selected for exploration}] \\
&= \Pr[\Delta_{A^0,A_{2,2}}^k \leq 0] \Pr[\text{selected for exploration}]
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|\text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|S_{1,1}^k]|\text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\text{explore}] \Pr[\text{explore}]
\end{aligned}$$

where the first equality is by the law of iterated expectation and the second equality is by definition of $\Delta_{A^0,A_{2,2}}^k$.

Therefore, we have:

$$\begin{aligned}
&\mathbb{E}[\mu_{A^0} - \mu_{A_{2,2}}|\text{explore}] \Pr[\text{explore}] \\
&= \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3] + \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_4] \Pr[\xi_4] \\
&= \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3] + \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\Delta_{A^0,A_{2,2}} < 0] \Pr[\Delta_{A^0,A_{2,2}}^k < 0] \cdot \frac{1}{L} \\
&= \left(1 - \frac{1}{L}\right) \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3] + \frac{1}{L} \cdot \mathbb{E}[\Delta_{A^0,A_{2,2}}^k] \\
&= \frac{L-1}{L} \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3] + \frac{1}{L} \cdot (\mu_{1,2}^{(0)} + \mu_{2,1}^{(0)} - 2\mu_{2,2}^{(0)})
\end{aligned}$$

The BIC condition can be written as:

$$\begin{aligned}
&\mathbb{E}[\mu_{A_{2,2}} - \mu_{A^0}|\neg\text{explore}] \Pr[\neg\text{explore}] \\
&= \mu_{2,2}^{(0)} - (\mu_{1,2}^{(0)} + \mu_{2,1}^{(0)}) + \frac{L-1}{L} \cdot \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3] + \frac{1}{L} \cdot ((\mu_{1,2}^{(0)} + \mu_{2,1}^{(0)}) - \mu_{2,2}^{(0)}) \\
&= \frac{L-1}{L} \left(\mu_{1,2}^{(0)} + \mu_{2,1}^{(0)} - \mu_{2,2}^{(0)} + \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3]\right)
\end{aligned}$$

Solving for L, we obtain the following condition:

$$L \geq 1 + \frac{2\epsilon}{\mu_{1,2}^{(0)} + \mu_{2,1}^{(0)} - \mu_{2,2}^{(0)} + \mathbb{E}[\Delta_{A^0,A_{2,2}}^k|\xi_3] \Pr[\xi_3] - 2\epsilon}$$

To ensure that this lower bound is not vacuous, we choose ϵ small enough such that the denominator is positive. $\square$

C.2 Accelerated Exploration Proofs

First, we state the following theorem from Jun et al. [2019] on finite sample error for low-rank bilinear bandits.

Theorem C.1 ([Jun et al., 2019]). *There exists a constant C such that for*

$$n_{\mathcal{P}} \geq C \cdot \sigma^2(g_0^2 + g_1^2) \cdot \frac{\kappa^6}{d\sigma^2_{\min}(K^*)} r(r + \log(d))$$

with probability at least $1 - 2/d_2^3$, we have

$$\left\| \hat{K} - K^* \right\|_F \leq C_1 \kappa^2 \sigma \sqrt{\frac{dr}{n_{\mathcal{P}}}} \tag{12}$$

where C_1 is an absolute constant, K^* is the mean reward matrix defined by $K^*_{i,j} = \mu_{i,j}$ with rank r, $\hat{K}$ is the noisy estimate of K^* using $n_{\mathcal{P}}$ samples of each atom, $\kappa = \sigma_{\max}(K^*)/\sigma_{\min}(K^*)$. Let $K^* = URV^\top$ be the SVD of K^*. Let (g_0, g_1) are the smallest values such that for all $i, j \in [d]$

$$\sum_{k=1}^{r} U_{ik}^2 \leq g_0 r/d \quad \sum_{k=1}^{r} V_{jk}^2 \leq g_0 r/d$$

$$\left| \sum_{k=1}^{r} U_{ik}(\sigma_k(K^*)/\sigma_{\max}(K^*))V_{jk} \right| \leq g_1 \sqrt{\frac{r}{d^2}}$$

Proof of Theorem 3.3. We begin by stating the formal theorem for accelerated exploration:

Theorem C.2 (Accelerated Exploration BIC). *Given $n_{\mathcal{P}}$ samples of all atoms where*

$$n_{cP} \geq \frac{N^6 C_1^2 \kappa^4 \sigma^2 dr}{4C_0^2(\Delta^{(t)}_{(b^{(t)})} - \epsilon N^2)^2}$$

the inverse gap weighting recommendation policy is two-sided ϵ-BIC. The total regret during this stage is $O(N\sqrt{dT\log(T)})$, which asymptotically matches the optimal regret of combinatorial semi-bandits.

Proof. We want to show that given a recommendation for any action $A \in \mathcal{A}$, the learner would not switch to some other action A'. Formally, we want to ensure the following condition:

$$\mathbb{E}[\mu_A - \mu_{A'} | \text{rec}^{(t)} = A] \Pr[\text{rec}^{(t)} = A] \geq \epsilon$$

Let $\Delta^{(t)}_{A,A'} = \mathbb{E}[\mu_A - \mu_{A'}|\mathcal{S}]$ denote the posterior gap between action A and A' given the data collected during the warm-start stage. Let $\Delta^{(t)}_A = \min_{A' \neq A} \Delta^{(t)}_{A,A'}$ denote the minimal posterior gap between action A and any other action. Then, when action A is recommended at time-step t, it means either 1) A is indeed the posterior best action at this time-step and $\Delta^{(t)}_A > 0$ or 2) A is not the posterior best action and $\Delta^{(t)}_A \geq 0$. We have

$$\begin{aligned}
&\mathbb{E}[\Delta^{(t)}_A | \text{rec}^{(t)} = A] \Pr[\text{rec}^{(t)} = A] \\
&= \mathbb{E}[\mathbb{E}[\mu_A - \max_{A' \in \mathcal{A}} \mu_{A'} | \mathcal{S}] | \text{rec}^{(t)} = A] \Pr[A^{(t)} = A] \\
&= \mathbb{E}[\mathbb{E}[\mu_A|\mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'}|\mathcal{S}] | A^{(t)} = b^{(t)}] \cdot \Pr[A^{(t)} = b^{(t)}] \\
&+ \mathbb{E}[\mathbb{E}[\mu_A|\mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'}|\mathcal{S}] | A^{(t)} \neq b^{(t)}] \cdot \Pr[A^{(t)} \neq b^{(t)}]
\end{aligned}$$

We proceed to analyze the lower bound for each case separately.

Exploitation: Recommended action $A^{(t)} = b^{(t)}$. By construction, the posterior best action is recommended with probability $p^{(t)}_{b^{(t)}} = 1 - \sum_{A \neq b^{(t)}} \frac{1}{N^2 + \gamma(\hat{r}^{(t)}_{b^{(t)}} - \hat{r}^{(t)}_A)}$. Since $\gamma > 0$, we observe that the probability of recommending any other action $A \neq b^{(t)}$ is at most $1/N^2$. Hence, we have $p^{(t)}_{b^{(t)}} \geq 1/N^2$. Therefore, we can write the reward gap in this case as:

$$\mathbb{E}[\mathbb{E}[\mu_A|\mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'}|\mathcal{S}] | A^{(t)} = b^{(t)}] \cdot \Pr[A^{(t)} = b^{(t)}] \geq \frac{1}{N^2} \cdot \Delta^{(t)}_{b^{(t)}}$$

Exploration: Recommended action $A^{(t)} \neq b^{(t)}$. The reward gap in this case can be written as follows.

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\mu_A|\mathcal{S}] - \max_{A' \in \mathcal{A}} \mathbb{E}[\mu_{A'}|\mathcal{S}]|A^{(t)} \neq b^{(t)}] \cdot \Pr[A^{(t)} \neq b^{(t)}] \\
&= \sum_{A \neq b^{(t)}} p_A^{(t)} \left(\mathbb{E}[\mu_A - \mu_{b^{(t)}}|\mathcal{S}]\right) \\
&= \sum_{A \neq b^{(t)}} \frac{1}{N^2 \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})} \left(\mathbb{E}[\mu_{b^{(t)}} - \mu_A|\mathcal{S}]\right) \\
&= -\mathbb{E}\left[\sum_{A \neq b^{(t)}} \frac{1}{\gamma} \cdot \frac{\gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}{N^2 + \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}\right] \\
&< -\frac{N^2 - 1}{\gamma} \qquad \left(\text{since } \frac{\gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})}{N^2 + \gamma(\hat{r}_{b^{(t)}}^{(t)} - \hat{r}_A^{(t)})} < 1\right) \\
&< -\frac{N^2}{\gamma}
\end{aligned}$$

Hence, for the BIC condition to hold, it suffices to show that

$$\begin{aligned}
&\mathbb{E}[\Delta_A^{(t)}|\mathrm{rec}^{(t)} = A] \Pr[\mathrm{rec}^{(t)} = A] \geq \epsilon \\
\iff\ & \frac{\Delta_{b^{(t)}}^{(t)}}{N^2} - \frac{N^2}{\gamma} \geq \epsilon \\
\iff\ & \gamma \geq \frac{N^4}{\Delta_{b^{(t)}}^{(t)} - \epsilon N^2}
\end{aligned}$$

By definition, we have $\gamma = C_0 \cdot N\sqrt{1/\phi^{(t)}}$. Then, combining with the condition above, we derive the requirement for the minimum prediction error at time-step t as:

$$\begin{aligned}
&\gamma \geq \frac{N^4}{\Delta_{b^{(t)}}^{(t)} - \epsilon N^2} \\
\iff\ & C_0 \cdot \frac{N}{\sqrt{\phi^{(t)}}} \geq \frac{N^4}{\Delta_{b^{(t)}}^{(t)} - \epsilon N^2} \\
\iff\ & \phi^{(t)} \leq \frac{C_0^2 \cdot (\Delta_{b^{(t)}}^{(t)} - \epsilon N^2)^2}{N^6}
\end{aligned}$$

Then, we use the theoretical guarantee of Theorem 2 in Jun et al. [2019] for bilinear bandits: Hence, it suffices to have

$$\begin{aligned}
\phi^{(t)} &\leq \frac{C_0^2 \cdot (\Delta_{b^{(t)}}^{(t)} - \epsilon N^2)^2}{N^6} \\
\frac{C_1^2 \kappa^4 \sigma^2 dr}{4 n_{\mathcal{P}}} &\leq \frac{C_0^2 \cdot (\Delta_{b^{(t)}}^{(t)} - \epsilon N^2)^2}{N^6} \\
n_{cP} &\geq \frac{N^6 C_1^2 \kappa^4 \sigma^2 dr}{4 C_0^2 (\Delta_{(b^{(t)})}^{(t)} - \epsilon N^2)^2}
\end{aligned}$$

Regret Analysis Following the analysis of Foster and Rakhlin [2020], with probability at least $1 - \delta$, the regret upper bound of the inverse gap weighting algorithm is $O(N\sqrt{T \cdot \phi^{(T)} \log(2/\delta)})$. □

D Experiment Detail

In this section, we provide the experimental details and analysis of Figure 1 that were previously omitted from the main body.

Experimental Details We consider a stylized setting with two types of agents and two types of arms as described in Section 2. All agents prefer to match with arms of type 1 and all arms prefer to match with agents of type 1. Our goal is to incentivize all agents and arms to explore all possible alternative matches and minimize regret.

We consider an online setting with a time horizon of $T = 20000$. At each time-step $t \in [T]$, 8 units arrive in a batch: two units for each type of agent or arm. The user profile for each agent of type 1 (resp. type 2) is $x_1^{(t)} = [v^{(t)}0]$ (resp. $x_2^{(t)} = [0v^{(t)}]$) where $v^{(t)} \sim \mathsf{Unif}[0,1]$. Similarly, the user profile for each arm of type 1 (resp. type 2) is $a_1^{(t)} = [u^{(t)}0]$ (resp. $a_2^{(t)}$) where $u^{(t)} \sim \mathsf{Unif}[0,1]$. The latent matrix Σ is generated as $\begin{pmatrix} 1 & 0.6 \\ 0.4 & 0.2 \end{pmatrix}$ to ensure that all agents prefer arms of type 1 and all arms prefer agents of type 1. Finally, the realized reward is generated by adding independent Gaussian noise $\eta_{i,j}^{(t)} \sim \mathcal{N}(0, 0.01)$ to each inner product of the user profiles.

Using Theorem 3.2, we calculate a lower bound on the phase length L of Algorithm 1 such that the ϵ-BIC condition (Definition 2.1) is satisfied for all agents and arms. Then, we calculate the number of samples needed to ensure that the efficient oracle in Section 3.2 is well-defined. We calculate the regret incurred by the combined algorithm by summing over the gap between the realized reward of the chosen action and the optimal matching at each time-step. This experiment is repeated 10 times and we report the regret and the standard error incurred at each time-step.

Results Our result is consistent with that of prior work in incentivized exploration. In the first stage of collecting 'warm-start' samples (Algorithm 1), we observe linear regret due to construction of the recommendation policy. Note that linear regret is also the state-of-the-art regret for the initial sample collection [Mansour et al., 2015]. When the second stage begins and we run the inverse gap weighting algorithm, the regret growth immediately decreases as the platform can explore more efficiently. In a real-life two-sided matching market, the platform can collect the initial samples by buying them, thus incurring no regret for the firs stage. Then, the platform only has to use the inverse gap weighting algorithm and observe sub-linear regret during its running time.

Future Work for experiments In our next revision, we aim to run more experiments to complement our theoretical results and explore how the regret changes in response to changes in hyperparameters. Particularly, we are interested in running experiments with more types of agents and arms, more number of agents and arms at each time-step, higher dimension of the user profiles, and varying gaps in the prior mean reward between different matches.