Multi-Player Approaches for Dueling Bandits

Or Raveh
The University of Tokyo
RIKEN AIP

Junya Honda
Kyoto University
RIKEN AIP

Masashi Sugiyama
RIKEN AIP
The University of Tokyo

Abstract

Fine-tuning large deep networks with preference-based human feedback has seen promising results. As user numbers grow and tasks shift to complex datasets like images or videos, distributed approaches become essential for efficiently gathering feedback. To address this, we introduce a multiplayer dueling bandit problem, highlighting that exploring non-informative candidate pairs becomes especially challenging in a collaborative environment. We demonstrate that the use of a Follow Your Leader black-box approach matches the asymptotic regret lower bound when utilizing known dueling bandit algorithms as a foundation. Additionally, we propose and analyze a message-passing fully distributed approach with a novel Condorcet-Winner recommendation protocol, resulting in expedited exploration in the nonasymptotic regime which reduces regret. Our experimental comparisons reveal that our multiplayer algorithms surpass single-player benchmark algorithms, underscoring their efficacy in addressing the nuanced challenges of this setting.

1 INTRODUCTION

Multi-armed bandit (MAB) (Auer et al., 2002) problems, which involve balancing exploration and exploitation to maximize rewards, are key in decision-making under uncertainty. Notable variations of MAB include the dueling-bandit problem and the cooperative multiplayer MAB problem. In the dueling-bandit scenario (Yue et al., 2012), feedback comes from pairwise comparisons between K arms. This is useful in human-feedback driven tasks such as ranker evaluation (Radlinski et al., 2008) and fine-tuning large language models (Rafailov et al., 2024). These works utilize preference-based optimization, which is based on ideas from dueling bandits or their contextual version (Nika et al., 2024; Yan et al., 2022). Approaches include optimistic algorithms (Zoghi et al., 2014) and algorithms that rely on information theoretic calculations (Komiyama et al., 2015). Meanwhile, the cooperative multiplayer MAB focuses on a group of M players collaboratively solving challenges in a distributed environment, enhancing learning through shared information. This is accomplished either by a leader-followers approach (Wang et al., 2020) or by a fully distributed approach (Dubey, 2020). Cooperative multiplayer MAB finds applications in fields like multi-robot systems (Lacerda et al., 2022) and distributed recommender systems (Rosaci et al., 2009).

As training deep networks with human feedback becomes more popular, user numbers grow and tasks shift into complex datasets like images or videos (Yang et al., 2024; Xu et al., 2024), which makes distributed approaches important for efficiently gathering feedback. A compelling application arises in fine-tuning large-scale distributed video generation systems (Ho et al., 2022), where a network of backend servers directs users to local algorithms that collect preference-based feedback regarding generated video quality. These systems, which must handle high-demand video file communication, optimize performance by relying on low-latency local interactions with nearby servers. Similarly, large-scale distributed ranker evaluation systems (Eirinaki et al., 2018) also benefit from collaboratively collecting preference-based information. Another example includes a network of self-ordering machines in regional fast-food chains, where each machine collects customer preferences for food products (Arsat et al., 2023). Local communication is beneficial due to varied preferences across different areas.

To address this, we introduce the M-player K-arm cooperative dueling bandit problem. This setting, unlike its multiplayer MAB counterpart, demands complex

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

coordination for exploration-exploitation across pairs of arms considering the quadratic number of available comparisons. In multiplayer MAB, communication delays can lead to drawing suboptimal arms, but these still provide useful information regarding the optimal one that reduces future regret. In contrast, in multiplayer dueling bandits, communication delays can lead to pulling pairs of either identical or non-identical suboptimal arms. This not only results in immediate regret but also provides no new information in the former case, and limited information in the latter compared to pulling a pair of optimal and suboptimal arms. Thus, a careful communication strategy that addresses the unique challenges of the multiplayer dueling bandit setting is crucial to achieving low-regret algorithms.

In this study, our focus centers on the widely explored Condorcet Winner (CW) assumption, in which there exists a unique preferable arm (Urvoy et al., 2013). We introduce a Follow Your Leader (FYL) black-box algorithm, that aligns well with existing dueling-bandit algorithms like Relative Upper Confidence Bound (RUCB) (Zoghi et al., 2014) and Relative Minimal Empirical Divergence (RMED) (Komiyama et al., 2015), and offers a more natural fit than multiplayer MAB variants (Landgren et al., 2018). Recognizing the limitations of relying on one leader for exploration (Yang et al., 2023), we also propose a decentralized extension to RUCB. Unlike in an MAB counterpart (Dubey, 2020), we demonstrate that incorporating an additional CW recommendation protocol between players significantly accelerates the identification of the CW in the nonasymptotic regime. We validate our approach with real-world data experiments. To summarize, our contributions are:

- We propose an intuitive black-box algorithm that can be integrated with existing single-player dueling bandit algorithms, conduct comprehensive regret analysis, and show it matches an established asymptotic lower bound when initialized properly.
- We conduct an in-depth regret analysis of a fully decentralized multiplayer algorithm based on RUCB, showing asymptotic optimality of the incurred regret up to constants.
- We demonstrate that a novel CW recommendation protocol results in a quick identification of the CW in the nonasymptotic regime.
- Through experiments on real data, we show that our algorithms exhibit superior performance compared to a single-player dueling bandit system.

The remainder of this paper is structured as follows: Section 2 introduces the problem and Section 3 discusses a regret lower bound. In Sections 4 and 5 we propose and analyze our algorithms. Section 6 showcases experimental results, and Section 7 concludes the paper. More related works are discussed in Appendix A, and Computational and space complexities as well as communication costs are discussed in Appendix B.

2 PROBLEM FORMULATION

For some matrix W, w_{ij} will denote the (i, j)-th element. In the M-player K-arm dueling bandit problem, each player $m \in [M] = \{1, 2, \ldots, M\}$ engages in the environment at each round by drawing an arm pair $(i_m(t), j_m(t)) \in [K]^2$ and receiving a Bernoulli reward feedback $r_{i_m(t),j_m(t)} = 1 - r_{j_m(t),i_m(t)}$, where $\mathbb{E}r_{i_m(t),j_m(t)} = q_{i_m(t),j_m(t)}$ for preference matrix $Q \in \mathbb{R}^{K\times K}$. Here, $\mathbb{E}X$ stands for the expected value of some random variable X. Q is a matrix unknown to players with $q_{ii} = 0.5$ and $q_{ij} + q_{ji} = 1$ for all $i, j \in [K]$. For brevity, we will use $r_m(t)$ when the context is clear. We assume there exists a Condorcet Winner (CW), which is an arm i such that $q_{ij} > 0.5$ for all $i \neq j$, and without loss of generality assume that the CW is arm 1. We also denote by $\mathcal{Q}_{\text{CW}}$ the class of preference matrices with a CW. The assumption regarding the existence of a CW remains widely adopted today. Notably, in Section 4 we utilize the RUCB algorithm (Zoghi et al., 2014) for single-player dueling bandits with a CW, which is based on the UCB approach for MAB, and achieves an order optimal $O(K \log T)$ regret with high probability. Additionally, the RMED algorithm (Komiyama et al., 2015) for the same setting has several versions which all employ an empirical divergence minimization approach. In Section 4 we use the RMED2FH version which has an $O(K \log T)$ regret and is known to be optimal.

Players are positioned on a connected undirected communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [M]$ denotes the vertices and $\mathcal{E}$ represents the set of edges. We denote $d(m, m')$ as the length of the shortest path between players m, m' on $\mathcal{G}$, and $D = \max_{m,m'} d(m, m')$ as the diameter of the graph. Additionally, the γ-power of $\mathcal{G}$ for a positive integer γ, denoted $\mathcal{G}_\gamma$, is a graph that shares $\mathcal{G}$'s vertices but includes an edge between any two vertices within a distance of γ or less in $\mathcal{G}$. The clique covering number of $\mathcal{G}$ is the minimum number of cliques required to include every vertex.

Communication operates under a message-passing protocol (Dubey, 2020; Madhushani et al., 2021) with a decay parameter γ. Each round, player m sends a message $x_m^{m'}(t)$ to each player m', which is received after $d(m, m') - 1$ rounds if $\gamma > d(m, m') - 1$ or is otherwise never obtained. This extends the traditional models of immediate reward sharing where only immediate neighbors communicate ($\gamma = 1$) (Kolla et al., 2018), and delayed communication in which all players communicate with each other under delays ($\gamma = D$) (Cesa-Bianchi et al., 2016). In multiplayer MAB, the two main approaches include distributed decision-making such as message passing UCB (Dubey, 2020), and a leader-follower paradigm. Notably, in the latter approach for Bernoulli rewards, Wang et al. (2020) introduced a distributed parsimonious exploration strategy with message passing akin to ours called DPE2, in which an elected leader explores using a UCB strategy, and followers exploit a designated arm sent by it. Their approach showcased an optimal regret bound, but required a meticulous derivation that cannot be applied in general.

Let $w_{ij}^m(t) = \sum_{\tau=1}^{t} \left(r_m(\tau)\overline{\zeta}_{i,j}^m(\tau) + (1-r_m)(\tau)\overline{\zeta}_{j,i}^m(\tau) \right)$ denote the number of wins of arm i over arm j experienced by player m until round t, where $\overline{\zeta}_{i,j}^m(t) = \mathbb{1}\{(i_m(t), j_m(t)) = (i,j)\}$ is an indicator for the draw of the ordered pair (i,j). Additionally, $N_{ij}^m(t) = \sum_{\tau=1}^{t}(w_{ij}^m(\tau) + w_{ji}^m(\tau))$ represents the number of visits to (i,j) or (j,i). We denote $\Delta_{ij} = q_{ij} - 1/2$ as the reward gap, and define the M-player cumulative regret as $\mathcal{R}(T) = \sum_{m=1}^{M}\sum_{t=1}^{T} 0.5(\Delta_{1 i_m(t)} + \Delta_{1 j_m(t)})$. Note that the instantaneous regret for a given player is zero only if the player draws the CW twice in a given round. Additionally, we define $\Delta_{1\max} := \max\{\Delta_{1i}\}$ and $\Delta_{1\min} := \min_{i>1}\{\Delta_{1i}\}$.

3 ASYMPTOTIC LOWER BOUND

We first discuss an asymptotic lower bound for the multiplayer dueling bandit setting in Theorem 3.1. Let $\mathcal{O}_i = \{j | j \in [K], q_{ij} < 1/2\}$ denote the superiors of arm i. An algorithm is considered consistent over the class $\mathcal{Q}_{\mathrm{CW}}$ if $\mathbb{E}\mathcal{R}(T) = o(T^p)$ for all $p > 0$ and $Q \in \mathcal{Q}_{\mathrm{CW}}$. Furthermore, we define $N_{ij}(T) = \sum_{m=1}^{M} N_{ij}^m(T)$ as the total number of visits to (i,j) or (j,i) in the system, and $\mathrm{KL}(p,q) = p\log\frac{p}{q} + (1-p)\frac{1-p}{1-q}$ as the Kullback-Leibler (KL) divergence between Bernoulli distributions.

Theorem 3.1. *For any consistent algorithm on $\mathcal{Q}_{\mathrm{CW}}$ and $Q \in \mathcal{Q}_{\mathrm{CW}}$, the group regret obeys,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\mathcal{R}(T)}{\log T} \geq \sum_{i\in[K]\setminus\{1\}} \min_{j\in\mathcal{O}_i} \frac{\Delta_{1i} + \Delta_{1j}}{2\mathrm{KL}(q_{ij}, 1/2)}.$$

A proof can be found in Appendix C. This bound scales as $\Omega(K\log T)$, similar to the single-player case (Komiyama et al., 2015). The lack of dependence on the number of players, M, is typical in multiplayer MAB settings (Yang et al., 2023; Wang et al., 2020) and indicates that, after sufficient time has passed, a multiplayer system can perform as efficiently as a single-player system drawing M pairs per round, regardless of distributed effects. In Sections 4 and 5, we establish matching upper bounds that show this lower bound is tight. However, for a finite horizon T, the regret of any algorithm cannot be independent of M (Martínez-Rubio et al., 2019), so the regret bounds for our algorithms include M and K dependent $o(\log T)$ terms. Developing tight lower bounds and corresponding algorithms for these non-leading terms remains an open problem in dueling bandits and multiplayer MAB.

Algorithm 1: Follow Your Leader Black Box (FYLBB)

```
Input: Time horizon $T$, number of arms $K$, number of players $M$, communication graph
  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, Base algorithm Alg0, Leader Election
  algorithm LEAlg, Election time $T_{\mathrm{LE}}$.
Initialize: $i(m) = 1$ for $m \in \mathcal{V}$.
for $t = 1, \ldots, T_{\mathrm{LE}}$ do
    for $m = 1, \ldots, M$ do
        Communicate according to LEAlg.
        Draw arms $(c(t), d(t))$ according to Alg0.
for $t = T_{\mathrm{LE}} + 1, \ldots, T$ do
    Leader $m_l$:
    Draw arms $(c(t), d(t))$ according to Alg0.
    Get CW candidate $\mathrm{cw}(t)$ from Alg0.
    if $\mathrm{cw}(t-1) \neq \mathrm{cw}(t)$ then
        Send $\mathrm{cw}(t)$ over the communication graph.
    Followers:
    for $m \in [M] \setminus \{m_l\}$ do
        if new arm $i$ arrives from leader then
            $i(m) \leftarrow i$.
        Draw pair $(i(m), i(m))$.
```

4 FOLLOW YOUR LEADER BLACK BOX ALGORITHM

We now present our first main contribution: a Follow Your Leader multiplayer algorithm capable of employing a single-player base dueling bandit algorithm as a black box. We begin with the following assumption regarding the base algorithm.

Assumption 4.1. Given a single-player history up to time t, the base algorithm outputs:

(a) A pair of arms $(c(t), d(t))$ to draw.

(b) A CW candidate $\mathrm{cw}(t)$, such that $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathrm{cw}(t) \neq 1\}\right] \leq f(K, T, Q)$.

Furthermore, it has a regret upper bound of $\mathbb{E}\mathcal{R}(T) \leq g(K, T, Q)$, where f, g denote some functions, and $f = o(g)$ with respect to T.

Assumption 4.1 utilizes a structure that several single player dueling bandit algorithms such as RUCB and RMED share. While they do not declare it, these algorithms maintain an internal estimate of the CW, and evaluate the number of rounds in which it is incorrect to be either a constant or $o(\log T)$. They then concentrate on comparing the CW with other arms, thereby reducing the overall regret by a factor of K. To maintain generality and avoid algorithm-specific notation, Assumption 4.1 formalizes this structure using the functions f, g and the estimate $\mathrm{cw}(t)$. For instance, in RUCB $\mathrm{cw}(t)$ can be taken as the hypothesized best arm $\mathcal{B}$ (when exists), and the analysis around the beginning of Theorem 4 in Zoghi et al. (2014) corresponds to the bound in Assumption 4.1(b).

Since we do not assume a leader is given as input, the algorithm starts with a distributed leader-election phase, utilizing some algorithm we denote as LEAlg. This is a well-studied problem with a range of efficient algorithms available (Santoro, 2006; Kutten et al., 2015), typically requiring minimal communication. We concentrate on deterministically identified leader election, where each player is assumed to be initialized with a unique integer identifier and communicates integer values (Casteigts et al., 2019). Utilizing leader-election algorithms without these assumptions is straightforward, but results in a higher communication cost (Mitra et al., 2022). LEAlg classifies each player as either “leader” or “non-leader”, with only one player holding the former designation, and is completed in T_{LE} time steps, independent of T. An illustrative example of such an algorithm can be found in Appendix D, where $T_{\mathrm{LE}} = D + 1$. During this initial phase, all players independently draw arms based on the base algorithm. For rounds $t = T_{\mathrm{LE}} + 1, \ldots, T$, the elected leader m_l continues to draw arms similarly without receiving feedback from other players, and in rounds where $\mathrm{cw}(t-1) \neq \mathrm{cw}(t)$, initiates a message containing the new $\mathrm{cw}(t)$. This is facilitated through a message-passing protocol with a decay parameter equivalent to the diameter of the communication graph. Simultaneously, each follower m engages in pure exploitation by drawing the arm pair $(i(m), i(m))$, where $i(m)$ is the arm last received from the leader.

While our setup resembles the DPE2 algorithm for multiplayer MAB (Wang et al., 2020), Algorithm 1 works with any base algorithm, unlike DPE2 which utilizes a specific exploration strategy that requires a meticulous derivation. In addition, unique to the dueling-bandit setting, followers are restricted to pure exploitation of the form $(\mathrm{cw}(t), \mathrm{cw}(t))$, meaning they gain no information about the preference matrix. Thanks to this property, it is relatively easy to adapt single-player dueling bandit algorithms to this framework without the need to carefully adjust the leader's exploration rate, by presenting a candidate arm $\mathrm{cw}(t)$, which is typically a crucial element in the original algorithm. Consequently, our algorithm achieves broader applicability.

Theorem 4.2. *Under Assumption 4.1, Algorithm 1 satisfies,*

$$\mathbb{E}\mathcal{R}(T) \leq g(K,T,Q) + M\Delta_{1\,\max} f(K,T,Q) \\ + M(T_{\mathrm{LE}} + 2D)\Delta_{1\,\max}\ .$$

Theorem 4.2 demonstrates that, in the asymptotic regime, the second and third terms are small compared to the first one, and the regret of Follow Your Leader Black Box (FYLBB) is dominated by the single-player regret bound g, thus inheriting the same asymptotic performance.

Proof Outline The proof involves decomposing the group regret resulting from the leader election phase, the one incurred when followers draw the currently recommended arm, the one when they draw a different arm due to delays, and the leader's regret itself. We establish that despite the difficulties that occur from the use of a general baseline algorithm and delays, Assumption 4.1 and the lack of follower feedback allow us to bound those terms by f and g in a corresponding single-player setting with the base algorithm. This enables us to prove a regret bound despite the generality of this setting. The detailed proof is available in Appendix D.

The construction in Theorem 4.2 enables us to seamlessly adapt single-player dueling bandit results to multiplayer scenarios and even attain asymptotically optimal algorithms.

Corollary 4.3. *For any $\epsilon > 0$ and $\alpha > 1$, the following holds for the regret bound of Algorithm 1 with leader election as specified by Algorithm 3.*

(a) For RUCB (Zoghi et al., 2014) as the base algorithm,

$$\mathbb{E}\mathcal{R}(T) = \sum_{i=2}^{K} \frac{4\alpha \log T}{\Delta_{1i}} + O\left(MD + M\frac{K^2 \log K}{\Delta_{1\,\min}^2}\right)$$

(b) For RMED2FH (Komiyama et al., 2015) as the base algorithm,

$$\mathbb{E}\mathcal{R}(T) = \sum_{i=2}^{K} \min_{j\in\mathcal{O}_i} \frac{\Delta_{1i} + \Delta_{1j}}{2\mathrm{KL}(q_{ij}, 1/2)} \log T \\ + O\left(M(D + K^{2+\epsilon}\log\log T) + \frac{K\log T}{\log\log T}\right)$$

By employing RMED2FH as the base algorithm, our approach achieves performance that matches the lower bound from Theorem 3.1 in the asymptotic regime. Utilizing RUCB yields an $O(K \log T)$ algorithm that remains independent of M under these conditions. Notably, for both algorithms, the candidate $\mathrm{cw}(t)$ corresponds to an arm already retained by the original algorithms. Lemma D.1 is used to prove Corollary 4.3 by providing the bounds f, g for RUCB and RMED2FH.

While Algorithm 1 can leverage any base algorithm satisfying Assumption 4.1 with comparable asymptotic guarantees, its empirical performance is suboptimal in certain scenarios, as demonstrated in Section 6. This arises from the lack of cooperation for exploration, as indicated by the nonasymptotic term $O\left(Mf(K,T,Q)\right)$, representing the regret incurred by the leader in identifying the CW. Moreover, an FYLBB approach is less suitable for systems where players enter and exit dynamically as it complicates leader election. These concerns motivate us to develop an expedited exploration approach in the next section.

5 A FULLY DISTRIBUTED APPROACH

While Algorithm 1 offers an intuitive black-box approach, exploration without cooperation can lead to significantly large finite-time regret, motivating us to introduce a distributed strategy. We demonstrate that a novel message-passing extension of RUCB in Algorithm 2 where players share rewards and update the CW candidate based on recommendations from other players, is an effective distributed approach for this setting. Unlike a naive message-passing extension (Madhushani et al., 2021) which would modify the RUCB algorithm such that data sent by other players is only utilized to construct improved confidence intervals, in Algorithm 2 players also share CW candidates. This fosters a more effective exploration that results in expedited identification of the CW by players. As a result, compared to a naive message-passing application, the incurred from drawing suboptimal arms is reduced. Detailed proofs for the claims in this section are found in Appendix E.

Algorithm 2: Message-Passing RUCB (MP-RUCB)

```
1  Input: Time horizon $T$, number of arms $K$, number of players $M$, $\alpha > 1.2$, communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, decay parameter $\gamma$.
2  Initialize: $\tilde{W}^m := (\tilde{w}^m_{ij}), \tilde{N}^m := (\tilde{N}^m_{ij}) \leftarrow 0_{K \times K}, \mathcal{B}^m, \mathcal{B}^m_{\mathrm{rec}}, x_m(t) \leftarrow \emptyset,\ \forall\ t \le 1,\ m \in \mathcal{V}$.
3  for $t = 1, 2, \ldots, T$ do
4      for $m = 1, \ldots, M$ do
5          $U^m := (u^m_{ij}) \leftarrow \frac{\tilde{W}^m}{\tilde{N}^m} + \sqrt{\frac{\alpha \log t}{\tilde{N}^m}}$.
6          $u^m_{ii} \leftarrow 0.5$ for $i \in [K]$.
7          $\mathcal{C}^m \leftarrow \{i \mid u^m_{ij} \ge 0.5\ \forall j \in [K]\}$.
8          $\mathcal{B}^m \leftarrow \mathcal{B}^m \cap \mathcal{C}^m$.
9          If $|\mathcal{C}^m| = 1$ then $\mathcal{B}^m \leftarrow \mathcal{C}^m$.
10         If $|\mathcal{B}^m_{\mathrm{rec}} \cap \mathcal{C}^m| \ge 1$ then set $\mathcal{B}^m$ to a random element from $\mathcal{B}^m_{\mathrm{rec}} \cap \mathcal{C}^m$.
11         if $\mathcal{C}^m = \emptyset$ then
12             pick $i_m(t)$ randomly from $[K]$.
13         else
14             if $\mathcal{B}^m \cap \mathcal{C}^m \ne \emptyset$ then
15                 pick $i_m(t)$ from $\mathcal{B}^m$.
16             else
17                 pick $i_m(t)$ randomly from $\mathcal{C}^m$.
18         $j_m(t) \leftarrow \mathrm{argmax}_j\, u^m_{ji}$.
19         Draw $(i_m(t), j_m(t))$ and obtain $r_m(t)$.
20         $x_m(t) \leftarrow \langle m, t, i_m(t), j_m(t), r_m(t) \rangle$.
21         Send $x_m(t)$ in message-passing.
22     for $m = 1, \ldots, M$ do
23         $x^{\mathrm{batch}}_m(t) \leftarrow \{x_{m'}(t') \mid d(m, m') = t - t' \le \gamma - 1\}$.
24         Update $\tilde{W}^m$ with samples in $x^{\mathrm{batch}}_m(t)$.
25         $\tilde{N}^m \leftarrow \tilde{W}^m + (\tilde{W}^m)^T$.
26         $\mathcal{B}^m_{\mathrm{rec}} \leftarrow \{i \mid (i, i) \in x_{m'}(t') \land d(m, m') = t - t' \le \gamma - 1\}$.
```

In lines 5–21 of Algorithm 2, each player follows a local version of RUCB. Computing the UCB terms u^m_{ij} for each arm pair (i,j) is done elementwise in line 5, where we define $x/0 = 1$ for any number x. To perform this calculation, player m maintains the matrix $\tilde{W}^m(t) := (\tilde{w}^m_{ij}(t))$. Here, $\tilde{w}^m_{ij}(t)$ represents information about the number of wins of i against j up to round t, known to the player from their own and other players' experiences up to that round. Similarly, $\tilde{N}^m(t) := (\tilde{N}^m_{ij}(t))$ represents the known number of visits up to round t. This is distinct from the local counters $w^m_{ij}(t)$ and $N^m_{ij}(t)$, which results in each player utilizing information gathered by other players in their decision-making. Additionally, each player maintains its own champion set $\mathcal{C}^m$ with arms likely to win against all other arms, a CW candidate set $\mathcal{B}^m$ that may either contain one candidate arm or be empty, and a set of recommended arms $\mathcal{B}^m_{\mathrm{rec}}$. The CW candidate set $\mathcal{B}^m$ is updated with a random element from $\mathcal{C}^m \cap \mathcal{B}^m_{\mathrm{rec}}$ whenever this set is non-empty, and is otherwise updated based on $\mathcal{C}^m$. In line 18, if there is a tie we draw $j_m(t) \ne i_m(t)$. Players communicate through a message-passing protocol with a decay parameter γ, as described in Section 2. For simplicity, we denote the message sent by player m at round t as $x_m(t)$ in line 20, which contains the drawn arms and the sampled reward. In lines 23–26, players receive messages and update $\tilde{W}^m$ and $\tilde{N}^m$ based on all available arms and reward samples. Moreover, whenever a new message includes an exploitation draw of the same arm twice, that arm is added to the recommended set $\mathcal{B}^m_{\text{rec}}$ for use in the next round.

We begin with a lemma that shows with high probability that for rounds larger than

$$C(\delta) := \left(4MK^2\left(3 + 2\log\left(d^\gamma_{\max} + 1\right)\right)/\delta\right)^{\frac{1}{1.7\alpha - 1.4}} \ ,$$

visitation numbers to arm pairs other than $(1,1)$ cannot be too large across the system.

Lemma 5.1. *For any $\alpha > 1.2$, $\delta > 0$ and $T_0 \geq C(\delta)$, define $N^{m,T_0}_{ij}(t)$ as the number of visitations of player m to (i,j) between rounds T_0 and t, and $\overline{\Delta}_{ij} := \min\{\Delta_{1i}, \Delta_{1j}\}$ for $i,j \neq 1$, $\overline{\Delta}_{1i} := \Delta_{1i}$. Additionally, for some $\gamma' \leq \gamma$ let $\mathcal{C}(\mathcal{G}_{\gamma'})$ be the set of all cliques in $\mathcal{G}_{\gamma'}$. Then, with probability $1-\delta$, $\sum_{m \in C} N^{m,T_0}_{ij}(t) \leq \left[4\alpha \log t/\overline{\Delta}^2_{ij} + (\gamma'+2)|C|\right] \mathbb{1}\{i \neq j\}$ holds for all rounds $t > T_0$, parameters $\gamma' \leq \gamma$, cliques $C \in \mathcal{C}(\mathcal{G}_{\gamma'})$ and pairs $(i,j) \neq (1,1)$.*

The establishment of this bound is pivotal, as it allows us to both bound the leading term of the group regret with a multiplicative factor smaller than M, and to bound the non-leading term in T corresponding to the regret incurred before identifying the CW. A naive analysis akin to the single-player case would lead to only bounding the local counters $N^{m,T_0}_{ij}(t)$, resulting in a regret that scales like $O(MK\log T)$. To avoid that, we capitalize on the shared observations among players to limit the total number of visits within any clique, and rely on the fact that the first arm is drawn from the player-specific champion set regardless of recommendations. Unlike existing multiplayer MAB methods (Madhushani et al., 2021), we demonstrate the visitation counter bound holds with high probability even for rounds in which the pair (i,j) is not drawn by any player, and for any clique in the graph to facilitate faster CW identification.

Before presenting the group regret bound for Algorithm 2, we introduce the following term.

$$\tilde{C} = \frac{1.7\alpha - 2.4}{1.7\alpha - 1.4}\left(4MK^2\left(3 + 2\log\left(d^\gamma_{\max} + 1\right)\right)\right)^{\frac{1}{1.7\alpha - 1.4}} \ .$$

Theorem 5.2. *For any $\delta > 0$ and $\gamma' \leq \gamma$, define $\mathcal{D} := \sum_{i<j} 4\alpha/\overline{\Delta}^2_{ij}$, $\Gamma(m,\gamma')$ as the size of the largest clique player m belongs to in $\mathcal{G}_{\gamma'}$ and*

$$\hat{C} := \Delta_{1\max} \sum_{m=1}^{M} \min_{\gamma' \leq \gamma}\left(K^2(\gamma'+2) + \frac{2\mathcal{D}}{\Gamma(m,\gamma')}\log 2\mathcal{D}\right) + \left(2\tilde{C} + K(3\gamma+2)\right)M\Delta_{1\max} \ .$$

Let $\chi(\mathcal{G}_\gamma)$ denote the clique covering number of $\mathcal{G}_\gamma$. Then for any $\alpha > 1.4$ and all times T,

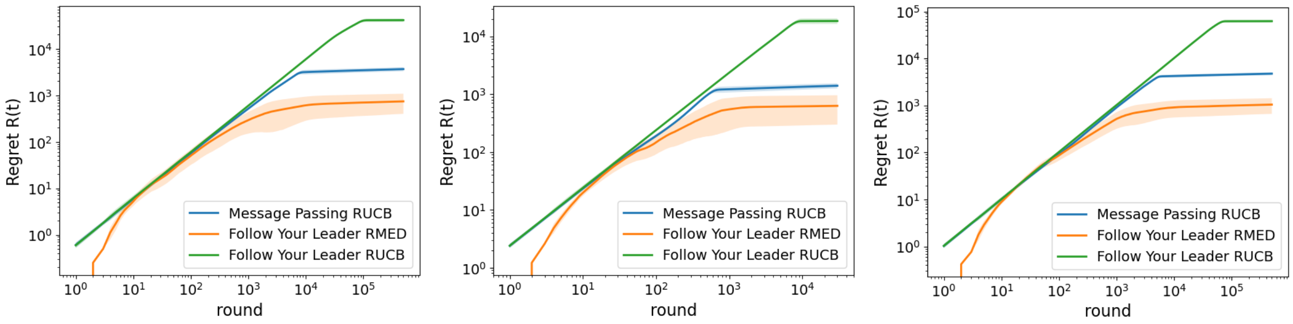
$$\mathbb{E}\mathcal{R}(T) \leq \sum_{j>1} \frac{2\alpha\chi(\mathcal{G}_\gamma)}{\Delta_{1j}}\log T + \hat{C} \ .$$

The regret bound presented in Theorem 5.2 demonstrates that a message-passing protocol is an effective distributed approach for the dueling bandit problem. It emphasizes the significance of leveraging candidate recommendations in reducing the regret incurred before identifying the CW, a characteristic unique to this scenario. In the asymptotic regime, the regret scales as $O\left(K\chi(\mathcal{G}_\gamma)\log T\right)$, consistent with distributed approaches in the multiplayer MAB setting (Dubey, 2020). For the no-communication case ($\gamma = 0$), this becomes an $O\left(MK\log T\right)$ bound, while for complete communication ($\gamma = D$), where every two players can communicate, the regret scales optimally as $O\left(K\log T\right)$.

Non-dominant Regret Term In the finite-time regime, the non-dominant term $\hat{C}$ is non-increasing with γ, and is determined separately for each graph $\mathcal{G}$ and parameter γ by minimization over two competing delay-dependent terms. While providing a direct expression for this term is difficult, in a complete communication setting, it scales like $O\left(MKD + \frac{K^2\log K}{\Delta^2_{1\min}} + MK^2\min\left\{\frac{\log K}{\Delta^2_{1\min}}, D\right\}\right)$, as discussed in Appendix E. Here, we observe a tradeoff between the delay D and the instance complexity $1/\Delta^2_{1\min}$. For challenging instances where $\Delta^2_{1\min}$ is sufficiently small, $\hat{C}$ scales as $O(K^2\log K/\Delta^2_{1\min})$ compared to the $O(MK^2\log K/\Delta^2_{1\min})$ non-leading term for FYL-RUCB. This observation helps explain the better empirical performance observed in Section 6.

Proof Outline Proving Theorem 5.2 is not straightforward, since it requires the establishment that each player actively identifies the CW quicker than a single-player counterpart, and that visitation counters within cliques cannot be too large with high probability. To achieve this, the proof relies on two parts. In the first one, we leverage Lemma 5.1 and demonstrate that, owing to CW candidate recommendations, each player identifies the CW within $O\left(\gamma + K^2\gamma' + K^2\log(K)/\Gamma(m,\gamma')\Delta^2_{1\min}\right)$ rounds with high probability for any clique such that $\gamma' \leq \gamma$. In the second part, we decompose the regret into a finite-time term corresponding to $\hat{C}$ and an asymptotic term, and bound them using the arbitrariness of γ' and the visitation count bound from Lemma 5.1. For the asymptotic term, each player consistently selects the CW as the first arm, leading to a high probability regret bound that is later converted into an expected regret bound.

Remark 5.3. The clique analysis in Theorem 5.2 is implicit, allowing the bound to depend on the optimal clique covering. In contrast, applying a similar clique analysis to Algorithm 1, where each clique has a designated leader, would necessitate actively finding the minimal clique covering, which is an NP-hard problem.

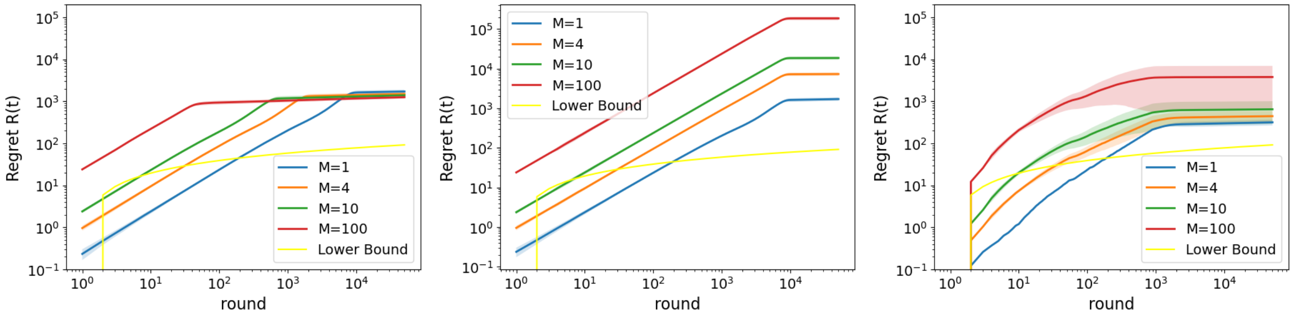


(a) Six Rankers

(b) Sushi

(c) Irish

Figure 1: Group regret for all datasets, using a complete communication graph with 10 players.



(a) Message Passing RUCB

(b) Follow Your Leader RUCB

(c) Follow Your Leader RMED

Figure 2: Group regret for the Sushi dataset: a complete graph with a varying number of players.

Additionally, using several leaders in this setting will result in an a regret that scales with the number of leaders, and not be asymptotically optimal.

6 EXPERIMENTS

In this section, we present experimental results to assess the empirical performance of Algorithms 1 and 2. The simulations utilized three preference matrices:

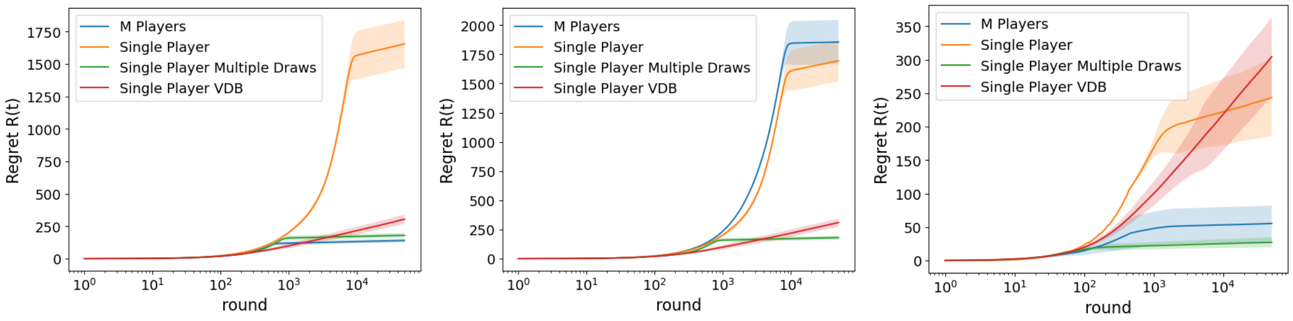
Six rankers: Used 6 arms, based on the six retrieval functions used in the engine of ArXiv.org (Yue and Joachims, 2011).

Sushi: Derived from the Sushi preference dataset (Kamishima, 2003), which contains preference data from 5000 users over 100 types of Sushi. We used a 10-kinds dataset (Mattei and Walsh, 2013) and converted it into a preference matrix, where q_{ij} denotes the ratio of users who prefer kind i over j.

Irish election: Based on the 2002 Dublin Meath election dataset (O'Neill, 2013), encompassing 64,081 votes over 14 candidates. We extracted the 10 most preferred candidates and constructed a preference matrix similarly to the Sushi dataset. This dataset was also used in Agarwal and Ghuge (2022).

The datasets above all feature a CW. We conducted a comparison between Algorithm 1 with both RMED2FH and RUCB as base algorithms, alongside Algorithm 2. Additionally, we evaluated the performance of the single-player Versatile-DB (VDB) algorithm (Saha and Gaillard, 2022), which is well-regarded for its strong theoretical guarantees and empirical effectiveness, making it a natural baseline for comparison with our multi-player algorithms. However, despite its effectiveness, this reduction-to-MAB approach does not track the CW, preventing it from satisfying Assumption 4.1, which is essential for extending it to a FYL multi-player setting. Moreover, implementing message passing for VDB presents additional challenges, as it requires communicating richer information about loss values, making its adaptation to the multiplayer setting more complex. Additional experimental results and details can be found in Appendix F.

Algorithmic Comparisons Across Datasets In Figure 1 we compare the algorithms across all datasets. The group regret per round is illustrated on a log-log scale, considering a complete communication graph with $M = 10$ nodes. For all datasets, we observe both the polynomial finite-time regime, dominated by exploration, and the linear asymptotic one, where the CW has been captured. In each dataset, FYLRMED

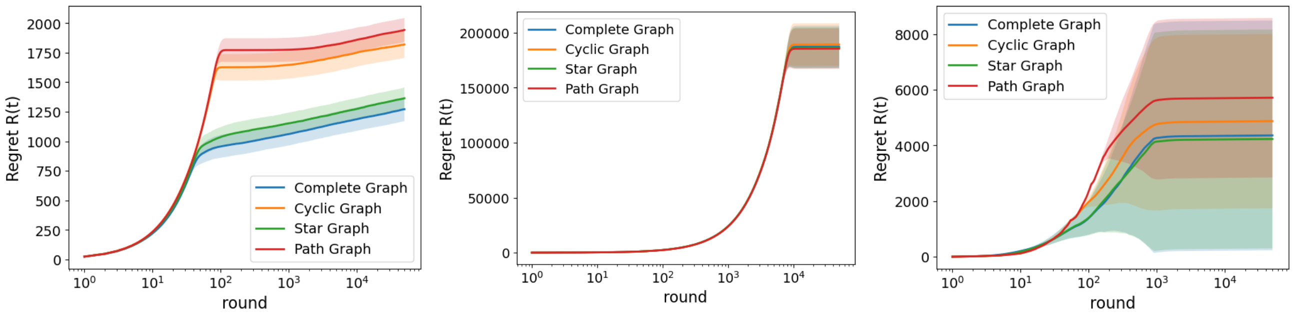


(a) Message Passing RUCB

(b) Follow Your Leader RUCB

(c) Follow Your Leader RMED

Figure 3: Average regret for the Sushi dataset with $M = 10$ players: a 10-player group, a single player, a single player with 10 draws in each round, and a single player VDB.



(a) Message Passing RUCB

(b) Follow Your Leader RUCB

(c) Follow Your Leader RMED

Figure 4: Group regret for the Sushi dataset: different communication graphs with $M = 100$ players.

consistently outperforms the other algorithms, while FYLRUCB is the least favorable. The superior performance of FYLRMED can be attributed to the enhanced exploration-exploitation tradeoff achieved by RMED2FH compared to RUCB, emphasizing that the performance of FYLBB is greatly influenced by the base algorithm. Message-passing RUCB showcases the significance of cooperation during exploration, leading to the early discovery of the CW compared to FYLRUCB, as indicated by the earlier transition into the asymptotic regime. Consequently, the regret is substantially lower, despite both algorithms using a similar decision-making process.

Experiments with a Varying Number of Players In Figure 2, we depict the group regret using the Sushi dataset on a complete communication graph with $M = 1, 4, 10, 100$, along with the lower bound. The results illustrate that, as anticipated by the regret bounds, the regret in the asymptotic regime remains unaffected by M. Message-Passing RUCB rapidly identifies the CW as the number of players increases due to collaborative exploration, leading to a regret that slightly reduces with M. For the two other algorithms only the leader explores, resulting in late identification of the CW and a regret that increases with the player count.

Comparisons with a Single Player In Figure 3, we show the average regret per player in a 10-player group for each algorithm. Additionally, we include the regret of a single player and the regret of a single player capable of making 10 actions in each round, divided by 10. For each figure, we also demonstrate the performance of a single player VDB algorithm with a learning rate of $4/\sqrt{t}$. We utilized the Sushi dataset and used a complete communication graph for the multiplayer setting. Notably, for all algorithms, a single player with only one action per round exhibits inferior performance in the asymptotic regime as implicated by its higher slope. Although less apparent given the time horizon of the experiment, this also holds for Figure 3b when T becomes large. Additionally, compared with VDB which incurs a significantly lower regret compared to RUCB and comparable to that of RMED under a single player setting, all multiplayer systems are superior to it both in slope and in the average regret.

These observations underscore the evident advantage of employing multiplayer systems. Conversely, a single player with 10 actions per round simulates a centralized 10-player system, where decision-making by a central entity benefits from all available information. Across all algorithms, the average player in a multiplayer system exhibits a comparable slope in the asymptotic regime to that of the average player in a centralized system, emphasizing their similar performance. This reaffirms that, asymptotically, our algorithms successfully emulate the behavior of a centralized system.

Comparisons Across Different Graph Structures Figure 4 depicts the group regret per round for the Sushi dataset and all algorithms, employing complete, cycle, star and path communication graphs with 100 players. For FYLBB algorithms, the leader is the central node in the star graph. Alternatively, it is the outermost node in the path graph. Notably, for Message Passing RUCB a complete graph and a star graph, with diameters of 1 and 2 respectively, outperform a cycle graph with a diameter of 50 and a path graph with a diameter of 99, due to delayed CW identification. This distinction is less apparent in the case of FYLBB algorithms, where the capture of the CW occurs later on average and dominates the incurred regret. The large variance in Figure 4c is discussed in Appendix F.

7 CONCLUSION AND FUTURE WORK

This paper introduced novel approaches to address the multiplayer dueling bandit problem, demonstrating the existence of efficient algorithms despite the increased complexity compared to an MAB setting. We first discussed a versatile black-box algorithm, leveraging either RUCB or RMED as base algorithms to achieve asymptotically optimal regret. However, as demonstrated in the experimental evaluation, finite-time performance can vary depending on the exploration efficiency of the base algorithm. To address this, we devised a novel message-passing protocol with CW recommendations for RUCB, showcasing more consistent performance and quick identification of the CW. All algorithms exhibited asymptotic regret comparable to that of a single player, a result validated theoretically and through simulations.

In future work, one promising avenue is the development of a black-box algorithm capable of compatibility with diverse base algorithms, while also enhancing collaborative exploration. A parallel approach involves devising a multiplayer black-box algorithm using a base multiplayer bandit algorithm, similar to approaches used in single-player dueling bandits (Saha and Gaillard, 2022). Additionally, making our algorithms more practical for real-world applications could involve incorporating federated learning (Wei et al., 2023) and privacy-preserving methods (Azize and Basu, 2022).

Acknowledgments

MS was supported by JST ASPIRE Grant Number JPMJAP2405 and by a grant from Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc. JH was supported by JSPS KAKENHI Grant Number JP21K11747.

References

Agarwal, A. and Ghuge, R. (2022). An asymptotically optimal batched algorithm for the dueling bandit problem. *Advances in Neural Information Processing Systems*, 35:28914–28927.

Agarwal, A., Ghuge, R., and Nagarajan, V. (2022). Batched dueling bandits. In *International Conference on Machine Learning*, pages 89–110. PMLR.

Agarwal, A., Johnson, N., and Agarwal, S. (2020). Choice bandits. *Advances in neural information processing systems*, 33:18399–18410.

Arsat, A., Hanafiah, M. H., and Che Shalifullizam, N. I. F. (2023). Fast-food restaurant consumer preferences in using self-service kiosks: An empirical assessment of the 4as marketing mix. *Journal of Culinary Science and Technology*, pages 1–12.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256.

Azize, A. and Basu, D. (2022). When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35:32199–32210.

Bistritz, I. and Leshem, A. (2018). Distributed multiplayer bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31.

Boursier, E. and Perchet, V. (2019). Sic-mmab: Synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems*, 32.

Boursier, E. and Perchet, V. (2024). A survey on multi-player bandits. *Journal of Machine Learning Research*, 25(137):1–45.

Casteigts, A., Métivier, Y., Robson, J. M., and Zemmari, A. (2019). Deterministic leader election takes Θ (D+log n) bit rounds. *Algorithmica*, 81:1901–1920.

Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. (2016). Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR.

Chawla, R., Sankararaman, A., Ganesh, A., and Shakkottai, S. (2020). The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3471–3481. PMLR.

Chen, B. and Frazier, P. I. (2017). Dueling bandits with weak regret. In *International Conference on Machine Learning*, pages 731–739. PMLR.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30:1–9.

Dubey, A. (2020). Cooperative multi-agent bandits with heavy tails. In *International Conference on Machine Learning*, pages 2730–2739. PMLR.

Eirinaki, M., Gao, J., Varlamis, I., and Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, 78:413–418.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. (2022). Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Howson, B., Filippi, S., and Pike-Burke, C. (2024). Quack: A multipurpose queuing algorithm for cooperative k-armed bandits. *arXiv preprint arXiv:2410.23867*.

Jamieson, K., Katariya, S., Deshpande, A., and Nowak, R. (2015). Sparse dueling bandits. In *Artificial Intelligence and Statistics*, pages 416–424. PMLR.

Kamishima, T. (2003). Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 583–588.

Kolla, R. K., Jagannathan, K., and Gopalan, A. (2018). Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795.

Kolpaczki, P., Bengs, V., and Hüllermeier, E. (2022). Non-stationary dueling bandits. *arXiv preprint arXiv:2202.00935*.

Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. (2015). Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on Learning Theory*, pages 1141–1154. PMLR.

Komiyama, J., Honda, J., and Nakagawa, H. (2016). Copeland dueling bandit problem: Regret lower bound, optimal algorithm, and computationally efficient algorithm. In *International Conference on Machine Learning*, pages 1235–1244. PMLR.

Kutten, S., Pandurangan, G., Peleg, D., Robinson, P., and Trehan, A. (2015). On the complexity of universal leader election. *Journal of the ACM (JACM)*, 62(1):1–27.

Lacerda, B., Gautier, A., Rutherford, A., Stephens, A., Street, C., and Hawes, N. (2022). Decision-making under uncertainty for multi-robot systems. *AI Communications*, 35(4):433–441.

Landgren, P., Srivastava, V., and Leonard, N. E. (2018). Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5239–5244. IEEE.

Landgren, P., Srivastava, V., and Leonard, N. E. (2021). Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445.

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

Madhushani, U., Dubey, A., Leonard, N., and Pentland, A. (2021). One more step towards reality: Cooperative bandits with imperfect communication. *Advances in Neural Information Processing Systems*, 34:7813–7824.

Martínez-Rubio, D., Kanade, V., and Rebeschini, P. (2019). Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32:1–12.

Mattei, N. and Walsh, T. (2013). Preflib: A library for preferences HTTP://WWW. PREFLIB.ORG. In *Proceedings of the Third International Conference on Algorithmic Decision Theory (ADT 2013)*, pages 259–270.

Mitola, J. and Maguire, G. Q. (1999). Cognitive radio: making software radios more personal. *IEEE personal communications*, 6(4):13–18.

Mitra, D., Cortesi, A., and Chaki, N. (2022). Alea: an anonymous leader election algorithm for synchronous distributed systems. In *Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference*, pages 46–58. Springer.

Nika, A., Mandal, D., Kamalaruban, P., Tzannetos, G., Radanović, G., and Singla, A. (2024). Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences. *arXiv preprint arXiv:2403.01857*.

O'Neill, J. (2013). Open stv. *http://www.OpenSTV.org*.

Radlinski, F., Kurup, M., and Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 43–52.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:1–14.

Rosaci, D., Sarne, G. M., and Garruzzo, S. (2009). Muaddib: A distributed recommender system supporting device adaptivity. *ACM Transactions on Information Systems (TOIS)*, 27(4):1–41.

Saha, A. and Asi, H. (2024). DP-Dueling: Learning from preference feedback without compromising user privacy. *arXiv preprint arXiv:2403.15045*.

Saha, A. and Gaillard, P. (2022). Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pages 19011–19026. PMLR.

Saha, A. and Gopalan, A. (2018). Battle of bandits. In *UAI*, pages 805–814.

Saha, A., Koren, T., and Mansour, Y. (2021). Adversarial dueling bandits. In *International Conference on Machine Learning*, pages 9235–9244. PMLR.

Sankararaman, A., Ganesh, A., and Shakkottai, S. (2019). Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35.

Santoro, N. (2006). *Design and analysis of distributed algorithms*. John Wiley & Sons.

Urvoy, T., Clerot, F., Féraud, R., and Naamane, S. (2013). Generic exploration and k-armed voting bandits. In *International Conference on Machine Learning*, pages 91–99. PMLR.

Wang, P. A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. (2020). Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR.

Wang, X. and Yang, L. (2023). Achieving near-optimal individual regret low communications in multi-agent bandits. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Wei, Z., Li, C., Xu, H., and Wang, H. (2023). Incentivized communication for federated bandits. *Advances in Neural Information Processing Systems*, 36:54399–54420.

Xia, F., Liu, H., Yue, Y., and Li, T. (2024). Beyond numeric awards: In-context dueling bandits with llm agents. *arXiv preprint arXiv:2407.01887*.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. (2024). Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Yan, X., Luo, C., Clarke, C. L., Craswell, N., Voorhees, E. M., and Castells, P. (2022). Human preferences as dueling bandits. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 567–577.

Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., and Li, X. (2024). Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951.

Yang, L., Wang, X., Hajiesmaili, M., Zhang, L., Lui, J., and Towsley, D. (2023). Cooperative multi-agent bandits: Distributed algorithms with optimal individual regret and constant communication costs. *arXiv preprint arXiv:2308.04314*.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2012). The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556.

Yue, Y. and Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208.

Yue, Y. and Joachims, T. (2011). Beat the mean bandit. In *Proceedings of the 28th Annual International Conference on Machine Learning*, pages 241–248.

Zoghi, M., Karnin, Z. S., Whiteson, S., and De Rijke, M. (2015). Copeland dueling bandits. *Advances in Neural Information Processing Systems*, 28:307–315.

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem. In *International Conference on Machine Learning*, pages 10–18. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]. The setting is explained in Section 2 and the algorithms are introduced in Sections 4 and 5.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]. An Analysis of the regret bounds can be found in Sections 4 and 5. A discussion about associated sample and computational complexities can be found in Appendix B.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]. See attached file.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes].

   (b) Complete proofs of all theoretical results. [Yes].

   (c) Clear explanations of any assumptions. [Yes].

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]. Code and instructions to reproduce the experimental results can be found in Appendix F.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]. Our experiments do not include training a neural network.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]. All figures include error bars that are calculated based on independent bandit experiments, as explained in Appendix F.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]. While bandit experiments are not heavy, general details about the computer resources are provided in Appendix F.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]. The datasets used for the experiments in the paper are referenced in Section 6 and Appendix F.

   (b) The license information of the assets, if applicable. [Yes]. We fully comply with the data usage policy detailed in Mattei and Walsh (2013).

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]. We do not use new assets.

   (d) Information about consent from data providers/curators. [Yes].We fully comply with the data usage policy detailed in Mattei and Walsh (2013).

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable].

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable].

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable].

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable].

A RELATED WORK

The dueling bandit problem, introduced as a useful variant of the MAB problem in Yue and Joachims (2009), initially underwent analysis with stringent assumptions about the preference matrix. These assumptions, including stochastic transitivity and stochastic triangle inequality, resulted in regret bounds of $O(K \log T)$ (Yue and Joachims, 2011; Yue et al., 2012). Subsequent works relaxed these assumptions by introducing the more realistic CW assumption (Zoghi et al., 2014; Komiyama et al., 2015; Chen and Frazier, 2017; Agarwal and Ghuge, 2022), which remains widely adopted today. Notably, the RUCB algorithm (Zoghi et al., 2014), based on the UCB approach for MAB, achieves an $O(K \log T)$ regret with high probability. Komiyama et al. (2015) established a comprehensive lower bound for this setting of the same order and provided the RMED family of algorithms that match it. Recently, Saha and Gaillard (2022) demonstrated a general reduction from dueling bandit algorithms to MAB and introduced a best-of-both-worlds algorithm effective for both adversarial and stochastic dueling bandits. Other recent advancements include a batched variant (Agarwal et al., 2022), a differentially private extension (Saha and Asi, 2024) and adversarial or non-stationary settings (Saha et al., 2021; Kolpaczki et al., 2022). Additionally, in cases where the CW does not exist, alternative definitions of winners, such as the Borda Winner (Jamieson et al., 2015) and the Copeland Winner (Komiyama et al., 2016; Zoghi et al., 2015), have been considered in dueling bandit literature. Further extensions to the original dueling bandit setting include systems in which more than two arms can be drawn at each round, such as battle of bandits (Saha and Gopalan, 2018) and choice bandits (Agarwal et al., 2020).

Considerable attention utilizing various approaches has been directed towards cooperative multiplayer MAB scenarios, revealing an $O(K \log T)$ regret lower bound (Dubey, 2020). Among them, in an approach called the leader-follower paradigm, players designate one or multiple leaders and base their arm selections on the leaders' instructions. In Kolla et al. (2018) and Landgren et al. (2018), followers emulate leaders employing a UCB strategy with neighbor-only communication, but the resulting algorithm does not match the lower bound. In the context of a Bernoulli bandit, Wang et al. (2020) introduced a distributed parsimonious exploration strategy with message passing akin to ours called DPE2, in which a leader explores using a UCB strategy, and followers exploit a designated arm sent by it. Their approach showcased a regret that aligns with the lower bound, but required a meticulous derivation that cannot be applied in general. Recently, Howson et al. (2024) suggested a general reduction approach from multiplayer to single player MAB, which can be applied to various settings including multiplayer dueling bandits. However, unlike our work which introduces both a leader-follower approach and a distributed one, their algorithm is only applicable when a leader can be elected, which is not always suitable as discussed in Section 4. In the fully distributed setting, which is more suitable in situations where players are non-synchronized or should have similar performance, both a message-passing approach (Dubey, 2020; Madhushani et al., 2021) and a running-consensus approach (Landgren et al., 2021) yield a regret that matches the lower bound up to constants. Alternatively, certain approaches adopt a gossiping strategy (Chawla et al., 2020; Sankararaman et al., 2019), in which a player can randomly communicate with one other player at each round.

A different approach to multiplayer MAB assumes that when multiple players select the same arm, a collision occurs, leading to a loss of information (Boursier and Perchet, 2024). This setting is primarily inspired by the cognitive radio problem, where smart devices must learn the optimal transmission configuration over a shared network (Mitola and Maguire, 1999). In this scenario, information loss arises when multiple users attempt to access the same channel simultaneously. The multiplayer MAB framework with collisions has received significant attention in recent years (Boursier and Perchet, 2019; Wang et al., 2020; Bistritz and Leshem, 2018), with various innovative approaches proposed to address the challenge. In contrast, our work is motivated by the task of fine-tuning large-scale networks with human feedback. This scenario is better captured by the non-collision cooperative setting, where collecting information about the same arm pair does not lead to interference.

Recently, fine-tuning large language models and text-to-image diffusion networks has proven effective and efficient through the use of preference-based human feedback (Christiano et al., 2017; Yang et al., 2024). Approaches include latent reward estimation and direct preference optimization, drawing inspiration from dueling bandit techniques (Xia et al., 2024; Nika et al., 2024)

B COMPUTATIONAL COMPLEXITY, SPACE COMPLEXITY AND COMMUNICATION COST

To efficiently handle message passing, where messages include the original player's id and the round they were created, players can maintain a table that records the ids of other players and the corresponding time it takes for messages to arrive from each one. This requires no prior knowledge, as the table can be updated when the first message from each player is received. Each player needs $O(M)$ space to store this table, enabling them to disregard redundant messages that may arrive via different paths in the communication graph. As each player can send up to γ messages per round (including forwarded messages), the computational complexity of message passing per player is $O(\gamma M)$, due to the need to filter out duplicate messages from neighboring players.

For Algorithm 1 initialized with the leader election algorithm from Algorithm 3, the space complexity per round during the leader election phase increases by an additive factor $O(M)$ relative to the base algorithm, while the computational complexity grows by $O(M^2)$ due to message passing. Once the leader is elected, it only initiates communication with its neighbors, so the space complexity remains the same as in the base algorithm, and the computational complexity per round increases by an additive factor of $O(M)$. Followers only need to store the currently recommended arm and handle message passing stemming from the leader, resulting in a space complexity of $O(M)$ and a per-round computational complexity of $O(M)$. In Algorithm 2, players maintain and update optimistic matrices of size K^2 in addition to handling message passing. As a result, the space complexity per player is $O(M + K^2)$, and the per-round computational complexity is $O(\gamma M + K^2)$.

In this work, we define the communication cost as the expected number of communication rounds, where a communication round is a round in which some player communicates. Since messages consist solely of integers, this measure aligns with other commonly used metrics, such as the total number of messages (Wang and Yang, 2023) and the total number of bits (Wang et al., 2020), differing only by a factor of $M\gamma$ and logarithmic terms. For Algorithm 1, the communication cost is bounded by $T_{\text{LE}} + 2f(K, T, Q)$, as shown in the proof for Theorem 4.2. Therefore, when $f(K, T, Q) = o(T)$, such as when using the base algorithms in Corollary 4.3, the communication cost is sub-logarithmic in T, as is often observed in multiplayer MAB (Yang et al., 2023) with a leader. By Lemma D.1, the communication cost of FYL-RUCB is independent of the horizon, while for FYL-RMED2FH it is $O(\log \log T)$. Similar to other distributed multiplayer MAB algorithms based on UCB (Landgren et al., 2021), the communication cost for Algorithm 2 is linear with T. Nevertheless, the computation cost per round remains low, as each message only contains a fixed number of integers. In comparison, running consensus algorithms (Landgren et al., 2021) in which each player shares the entire preference estimation matrix incurs a computation cost of $O(MK^2)$. Additionally, these algorithms either need to send non-integer numbers in the form of the preference matrix, or matrices with integers of the order $O(T)$ in the form of total visitation counters and arm comparison counters.

C LOWER BOUND PROOF

This section consists of two subsections. In Section C.1, we establish a precise definition for a canonical probability space tailored to the multiplayer dueling bandit scenario. Following this foundational definition, Section C.2 presents a comprehensive proof for the lower bound encapsulated in Theorem 3.1. To ensure clarity and coherence, we revisit key definitions introduced in Sections 2 and 3.

C.1 A Canonical Probability Space

Definition C.1. A stochastic dueling bandit environment is a collection of distributions $\nu = (P_{ij} : i, j \in [K])$, where K is the number of available actions, and P_{ij} is a Bernoulli distribution with mean q_{ij}, such that $q_{ij} = 1 - q_{ji}$.

In this context, the preference matrix $Q = (q_{ij}) \in \mathbb{R}^{K \times K}$ represents a given environment. Next, we define a class of preference matrices.

Definition C.2. A dueling bandit environment class is $\mathcal{Q} = (Q)$ is some collection of preference matrices. In particular, we define,

(a) The class of preference matrices with total ordering: $\mathcal{Q}_{\text{o}} = \{Q \mid i \prec j \Leftrightarrow q_{ij} < 1/2\}$.

(b) The class of preference matrices with a Condorcet Winner (CW): $\mathcal{Q}_{\text{CW}} = \{Q \mid \exists i : q_{ij} > 1/2\ \forall j \neq i\}$.

Note that $\mathcal{Q}_{\text{o}} \subset \mathcal{Q}_{\text{CW}}$. Now, let T be the time horizon. We consider M players communicating via a connected, undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Communication is bidirectional, and any message sent from player m may be obtained by player m' after $d(m, m') - 1$ rounds of the bandit problem, where $d(m, m')$ denotes the length of the shortest path between players m and m' on $\mathcal{G}$. Let $(i_m(t), j_m(t))$ denote the pair drawn by player m at time t, and $r_m(t)$ denote the corresponding reward. The power graph of order γ of $\mathcal{G}$, denoted by $\mathcal{G}_\gamma$, is defined to include an edge (m, m') if there exists a path of length at most γ in $\mathcal{G}$ between players m and m'. the neighborhood of m in $\mathcal{G}_\gamma$ is given by $\mathcal{N}_\gamma(m)$, including player m itself. We establish a communication protocol as follows.

Assumption C.3. Players can communicate in the following manner:

- Any player m is capable of sending a message $x_m^{m'}(t)$ to any other player $m' \in \mathcal{N}_\gamma(m)$ at time t, and this message may be received by player m' at time $t + \min(0, d(m, m') - 1)$. This is done via a message-passing protocol as described in Section 2
- The message $x_m^{m'}(t)$ is a function of the arms-reward triplets of player m up to and including time t, i.e. $x_m^{m'}(t) = F_{m,t}^{m'}\left(i_m(1), j_m(1), r_m(1), \ldots, i_m(t), j_m(t), r_m(t)\right)$ for any deterministic Borel function $F_{m,t}^{m'} : \mathbb{R}^{3t} \to \mathbb{R}^L$, with L being a nonnegative integer.

To define a measurable probability space for this setting, let us treat the players as ordered, such that player 1 is the first one and player M is the last. We define the outcome until time t and player m as the ordered set $H_{m,t} = (i_1(1), j_1(1), r_1(1), \ldots, i_m(t), j_m(t), r_m(t))$, the outcome space as $\Omega_{m,t} = ([K]^2 \times \mathbb{R})^{M(t-1)+mt}$, and the corresponding sigma-algebra as $\mathcal{F}_{m,t} = \mathcal{B}(\Omega_{m,t})$. Here, $\mathcal{B}(\cdot)$ stands for the Borel sigma-algebra. Using these, we define a measurable space for the entire multiplayer dueling bandit problem as $(\Omega_{M,T}, \mathcal{F}_{M,T})$, with the random variables of the drawn arms and obtained rewards as

$$
\begin{aligned}
i_m(t)\left(i_{1,1}, j_{1,1}, x_{1,1}, \ldots, i_{M,T}, j_{M,T}, r_{M,T}\right) &= i_{m,t}\ , \\
j_m(t)\left(i_{1,1}, j_{1,1}, x_{1,1}, \ldots, i_{M,T}, j_{M,T}, r_{M,T}\right) &= j_{m,t}\ , \\
r_m(t)\left(i_{1,1}, j_{1,1}, x_{1,1}, \ldots, i_{M,T}, j_{M,T}, r_{M,T}\right) &= r_{m,t}.
\end{aligned}
$$

For clerity, we also define the player-specific measurable space as $(\Omega_T^m, \mathcal{F}_T^m)$, where $H_t^m = (i_m(1), j_m(1), r_m(1), \ldots, i_m(t), j_m(t), r_m(t))$, $\Omega_t^m = \left([K]^2 \times \mathbb{R}\right)^t$, and $\mathcal{F}_t^m = \mathcal{B}(\Omega_t^m)$. Next, we define the multiplayer dueling bandit policy, which can be separated into a communication policy and an action policy. This separation is motivated by the fact that both aspects are under the control of the learner.

Definition C.4. A communication policy $\tilde{\Pi}_{\text{comm},\gamma}$ for the multiplayer dueling bandit setup is the sequence of messages $(F_{m,t}^{m'})_{m,m',t}$ from Assumption C.3, where $m \in [M], t \in [T]$ and $m' \in \mathcal{N}_\gamma(m)\ \forall m \in [M]$.

Definition C.5. An action policy $\tilde{\Pi}$ for the multiplayer dueling bandit setup is a sequence $(\tilde{\Pi}_t)_{t=1}^T$ where $\tilde{\Pi}_t = (\tilde{\Pi}_t^m)$, and $\tilde{\Pi}_t^m$ is a probability kernel from $\left(\Omega_{t-1}^m \times \mathbb{R}^{L_{m,t}^\gamma}, \mathcal{B}\left(\Omega_{t-1}^m \times \mathbb{R}^{L_{m,t}^\gamma}\right)\right)$ to $([K]^2, 2^{[K]^2})$, where $L_{m,t}^\gamma = L \sum_{m' \in \mathcal{N}_\gamma(m) \setminus \{m\}} (t - d(m, m'))$. We define $\tilde{\pi}_t = (\tilde{\pi}_t^m)$ as the corresponding density with respect to the counting measure on $([K]^2, 2^{[K]^2})$, so that $\forall i, j \in [K]$,

$$
\tilde{\pi}_t^m\left(i, j | h_{t-1}^m, z_t^m\right) = \tilde{\Pi}_t^m\left(\{i, j\} | h_{t-1}^m, z_t^m\right).
$$

Here, where z_t^m stands for all the messages received by player m until round t.

The communication policy $\tilde{\Pi}_{\text{comm},\gamma}$ and the action policy $\tilde{\Pi}$ as defined above illustrate the nature through which a learner operates, are easy to extract from an algorithm, and will help define the probability density function of the outcome. However, we can combine them into a more convenient policy definition that only takes into account the outcomes in $\Omega_{M,T}$.

Definition C.6. A policy Π for the multiplayer dueling bandit setup is a sequence $(\Pi_t)_{t=1}^T$ where $\Pi_t = (\Pi_t^m)$, and Π_t^m is a probability kernel from $\left(\Omega_{t-1}^m \times \prod_{m' \in \mathcal{N}_\gamma(m) \setminus \{m\}} \Omega_{t-d(m,m')}^{m'}, \mathcal{B}\left(\Omega_{t-1}^m \times \prod_{m' \in \mathcal{N}_\gamma(m) \setminus \{m\}} \Omega_{t-d(m,m')}^{m'}\right)\right)$ to $([K]^2, 2^{[K]^2})$. We define $\pi_t = (\pi_t^m)$ as the corresponding density with respect to the counting measure on $([K]^2, 2^{[K]^2})$, so that$\forall i, j \in [K]$

$$
\pi_t^m\left(i, j \middle| h_{t-1}^m \times \prod_{m' \in \mathcal{N}_\gamma(m) \setminus \{m\}} h_{t-d(m,m')}^{m'}\right) = \tilde{\pi}_t^m\left(i, j | h_{t-1}^m, z_t^m\right).
$$

We are now ready to define the probability measure on $(\Omega_{M,T}, \mathcal{F}_{M,T})$, which can be done in two ways similar to the MAB setting (Lattimore and Szepesvári, 2020). As a first definition, we require that -

- The conditional distribution of $i_m(t), j_m(t)$ given $i_1(1), \dots, r_{m-1}(t)$ (in case $m > 1$, or $i_1(1), \dots, r_M(t-1)$ otherwise) is Π_t^m almost surely.
- The conditional distribution of $r_m(t)$ given $i_1(1), \dots, r_{m-1}(t), i_m(t), j_m(t)$ is $P_{i_m(t), j_m(t)}(r_m(t))$ almost surely.

This defines a unique probability measure on $(\Omega_{M,T}, \mathcal{F}_{M,T})$. Alternatively, as a second definition, we can also define a probability measure with respect to a probability density function (pdf). Let λ be a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for which P_{ij} is absolutely continuous with respect to λ for all $i, j \in [K]$. denote $p_{ij} = dP_{ij}/d\lambda$, and let ρ be the counting measure on $([K]^2, 2^{[K]^2})$ (since P_{ij} is Bernoulli, we can take λ to be the Lebesgue measure, so that $p_{ij}(x) = q_{ij}x + (1 - q_{ij})x$). Utilizing Fubini's theorem and the properties of the Radon-Nikodym derivative, the pdf of an outcome is,

$$\begin{aligned} & p_{\nu\pi}(i_{1,1}, j_{1,1}, r_{1,1}, \dots, i_{M,T}, j_{M,T}, r_{M,T}) \\ & = \prod_{t=1}^{T} \prod_{m=1}^{M} \pi_t^m \left(i_{m,t}, j_{m,t} | h_{t-1}^m \times \prod_{m' \in \mathcal{N}_\gamma(m) \setminus \{m\}} h_{t-d(m,m')}^{m'} \right) p_{i_{m,t}, j_{m,t}}(r_{m,t}), \end{aligned}$$

and the measure can be calculated as,

$$\begin{aligned} \mathbb{P}_{\nu\pi}(B) = \int_B & \left(p_{\nu\pi}(i_{1,1}, j_{1,1}, r_{1,1}, \dots, i_{M,T}, j_{M,T}, r_{M,T})(\rho \times \lambda)^{MT} \right. \\ & \left. \times d(i_{1,1}, j_{1,1}, r_{1,1}, \dots, i_{M,T}, j_{M,T}, r_{M,T})\right) \ \forall B \in \mathcal{F}_{M,T}. \end{aligned}$$

As mentioned before, these two definitions result in the same probability measure.

C.2 Proof for Theorem 3.1

First, we present a useful decomposition for the group regret within the context of multiplayer dueling bandits. Throughout this section, we will frequently employ the notation $\mathbb{E}\mathcal{R}_{\mathcal{G},T}(\pi, \nu)$ instead of $\mathbb{E}\mathcal{R}(T)$ to underscore its dependence on various terms.

$$\begin{aligned} \mathbb{E}\mathcal{R}_{\mathcal{G},T}(\pi, \nu) &= \mathbb{E}_\nu \left[\sum_{t=1}^{T} \sum_{m=1}^{M} \frac{\Delta_{1 i_m(t)} + \Delta_{1 j_m(t)}}{2} \right] \\ &= \frac{1}{2} \sum_{i \in [K]} \sum_{j \in [K] \setminus \{i\}} \frac{\Delta_{1i} + \Delta_{1j}}{2} \mathbb{E}_\nu N_{ij}(T) + \sum_{i \in [K]} \Delta_{1i} \mathbb{E}_\nu N_{ii}(T). \end{aligned}$$

Here, the subscript ν signifies that the expectation is evaluated for bandit environment ν. In adittion, $N_{ij}(T) = \sum_{t=1}^{T} \sum_{m=1}^{M} (\mathbb{1}\{i_m(t) = i, j_m(t) = j\} + \mathbb{1}\{i_m(t) = j, j_m(t) = i\})$ for $i \neq j$ and $N_{ii}(T) = \sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{1}\{i_m(t) = i, j_m(t) = i\}\}$ represent the total number of visits to arms (i, j) by time T without order. We proceed to define the class of consistent policies, a class for which the lower bound holds.

Definition C.7. A policy π for the multiplayer dueling bandit setup is called consistent over a class of bandits $\mathcal{Q}$ if for all $Q \in \mathcal{Q}$ and $p > 0$, it holds that,

$$\lim_{T \to \infty} \frac{\mathbb{E}\mathcal{R}_{\mathcal{G},T}(\pi, \nu)}{T^p} = 0.$$

The class of consistent policies over $\mathcal{Q}$ is denoted by $\Pi_{\text{cons}}(\mathcal{Q})$.

Next, we present a divergence decomposition lemma tailored to the multiplayer dueling bandit setting. This will be used to prove the lower bound.

Lemma C.8. *Let $\nu = (P_{ij})$ and $\nu' = (P_{ij}^{'})$ be two stochastic dueling bandit environments. Fix some multiplayer policy π such that Assumption C.3 holds, and let $\mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'\pi}$ denote the probability measures for the canonical bandit models induced by the T-round interconnection of π and ν, ν' respectively. Then,*

$$\text{KL}\left(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi}\right) = \frac{1}{2} \sum_{i \in [K]} \sum_{j \in [K] \setminus \{i\}} \mathbb{E}_\nu[N_{ij}(T)] \text{KL}(q_{ij}, q_{ij}^{'}),$$

where $\mathrm{KL}(\mathbb{P},\mathbb{Q})$ *stands for the KL divergence between probability measures* $\mathbb{P},\mathbb{Q}$

$$\mathrm{KL}\left(\mathbb{P},\mathbb{Q}\right)=\mathbb{E}_{\mathbb{P}}\left[\log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right],$$

and $\mathrm{KL}(p,q)$ *stands for the KL divergence between two Bernoulli measures with means* p,q *respectively.*

Proof. Assume that $\mathrm{KL}\left(q_{ij},q_{ij}^{'}\right)<\infty$ for all $i,j\in[K]$. It follows from the KL-divergence definition that $P_{ij}\ll P_{ij}^{'}$. Next, define the measure $\lambda=\sum_{i=1}^{K}(P_{ij}+P_{ij}^{'})$ for which $P_{ij},P_{ij}^{'}\ll\lambda$ for all $i,j\in[K]$.

It is evident that the policy terms π_t^m remain consistent across bandit environments ν and bandit ν' when our interest lies in the same outcome. This consistency arises from Assumption C.3, which ensures that messages are deterministic functions of the outcome and are otherwise independent of the environment. Additionally, a policy is uniquely determined by action-reward pairs and messages. Therefore, using the pdf of the outcome with respect to the product measure $(\rho\times\lambda)^{MT}$ for each environment,

$$\log\frac{d\mathbb{P}_{\nu\pi}}{d\mathbb{P}_{\nu'\pi}}(i_{1,1},\dots,r_{M,T})=\sum_{t=1}^{T}\sum_{m=1}^{M}\log\frac{p_{i_{m,t},j_{m,t}}(r_{m,t})}{p_{i_{m,t},j_{m,t}}^{'}(r_{m,t})},$$

and

$$\mathbb{E}_{\nu}\left[\log\frac{d\mathbb{P}_{\nu\pi}}{d\mathbb{P}_{\nu'\pi}}(i_{1,1},\dots,r_{M,T})\right]=\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}_{\nu}\left[\log\frac{p_{i_{m,t},j_{m,t}}(r_{m,t})}{p_{i_{m,t},j_{m,t}}^{'}(r_{m,t})}\right].$$

For every term in the sum we have that,

$$\begin{aligned}\mathbb{E}_{\nu}\left[\log\frac{p_{i_{m,t},j_{m,t}}(r_{m,t})}{p_{i_{m,t},j_{m,t}}^{'}(r_{m,t})}\right]&=\mathbb{E}_{\nu}\left[\mathbb{E}_{\nu}\left[\log\frac{p_{i_{m,t},j_{m,t}}(r_{m,t})}{p_{i_{m,t},j_{m,t}}^{'}(r_{m,t})}\Big|i_{m,t},j_{m,t}\right]\right]\\&=\mathbb{E}_{\nu}\left[\mathrm{KL}(P_{i_{m,t},j_{m,t}},P_{i_{m,t},j_{m,t}}^{'})\right].\end{aligned}$$

Returning to the previous equation,

$$\begin{aligned}\mathbb{E}_{\nu}\left[\log\frac{d\mathbb{P}_{\nu\pi}}{d\mathbb{P}_{\nu'\pi}}(i_{1,1},\dots,r_{M,T})\right]&=\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}_{\nu}\left[\mathrm{KL}(P_{i_{m,t},j_{m,t}},P_{i_{m,t},j_{m,t}}^{'})\right]\\&=\sum_{i,j=1}^{K}\mathbb{E}_{\nu}\left[\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{1}\{(i_{m,t}=i,j_{m,t}=j)\}\mathrm{KL}(P_{ij},P_{ij}^{'})\right]\\&=\frac{1}{2}\sum_{i\in[K]}\sum_{j\in[K]\setminus\{i\}}\mathbb{E}_{\nu}[N_{ij}(T)]\mathrm{KL}(P_{ij},P_{ij}^{'})+\sum_{i\in[K]}\mathbb{E}_{\nu}[N_{ii}(T)]\mathrm{KL}(P_{ij},P_{ij}^{'})\\&=\frac{1}{2}\sum_{i\in[K]}\sum_{j\in[K]\setminus\{i\}}\mathbb{E}_{\nu}[N_{ij}(T)]\mathrm{KL}(q_{ij},q_{ij}^{'})+\sum_{i\in[K]}\mathbb{E}_{\nu}[N_{ii}(T)]\mathrm{KL}(q_{ii},q_{ii}^{'})\\&=\frac{1}{2}\sum_{i\in[K]}\sum_{j\in[K]\setminus\{i\}}\mathbb{E}_{\nu}[N_{ij}(T)]\mathrm{KL}(q_{ij},q_{ij}^{'}),\end{aligned}$$

where in the last transition we used the fact that $q_{ii}=q_{ii}^{'}=1/2$. Given that $d\mathbb{P}_{\nu\pi}/d\mathbb{P}_{\nu'\pi}<\infty$, it holds that,

$$\mathrm{KL}\left(\mathbb{P}_{\nu\pi},\mathbb{P}_{\nu'\pi}\right)=\mathbb{E}_{\nu}\left[\log\frac{d\mathbb{P}_{\nu\pi}}{d\mathbb{P}_{\nu'\pi}}\right].$$

It is evident that in our case the expectation is finite, so

$$\mathrm{KL}\left(\mathbb{P}_{\nu\pi},\mathbb{P}_{\nu'\pi}\right)=\frac{1}{2}\sum_{i\in[K]}\sum_{j\in[K]\setminus\{i\}}\mathbb{E}_{\nu}[N_{ij}(T)]\mathrm{KL}(q_{ij},q_{ij}^{'}).$$

In case $\mathrm{KL}(q_{ij},q_{ij}^{'})=\infty$ for some $i,j\in[K]$, this relation still holds since both sides of the equation become infinite. □

We also use the following lemma.

Lemma C.9. *For any consistent algorithm on $\mathcal{Q}_{\text{CW}}$ and $Q \in \mathcal{Q}_{\text{CW}}$, and for any arm $i \in [K] \setminus \{1\}$ the following holds.*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\sum_{j\in\mathcal{O}_i} \text{KL}(q_{ij}, 1/2) N_{ij}(T)\right]}{\log T} \geq 1.$$

Proof. Fix some arm $i \in [K] \setminus \{1\}$ and define $\mathcal{O}_i^{'} = \{j \mid j \in [K], q_{ij} \leq 1/2\}$. We consider a challenging instance $Q^{'}$, which is defined such that the means $q_{ij}^{'}$ for $j \in \mathcal{O}_i^{'}$ satisfy $\text{KL}(q_{ij}, q_{ij}^{'}) = \text{KL}(q_{ij}, 1/2) + \epsilon$ with $q_{ij}^{'} > 1/2$, for some $\epsilon > 0$ (we leave $q_{ii}^{'} = q_{ii} = 1/2$). The rest of the means remain the same as q_{ij}. For this new instance, $Q^{'} \in \mathcal{Q}_{\text{CW}}$ and arm i is the CW.

It is important to highlight that a proof scheme akin to ours that is used for MAB (Lattimore and Szepesvári, 2020) is applicable only to an unstructured class of bandits. This is due to the necessity of the unstructured assumption to construct the challenging instance $\nu^{'}$. The unstructured assumption becomes crucial when attempting to change the mean of a single arm, as the presence of interdependence among arms restricts such modifications. In the dueling bandit scenario, we can directly provide a challenging instance within the structured class, even though the arms' means may be correlated. This distinction allows for a more flexible approach to constructing challenging instances within the dueling bandit framework.

Note that since only the means q_{ij} for $j \in \mathcal{O}_i^{'}$ are different in the two environments, Lemma C.8 states that,

$$\text{KL}\left(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi}\right) = \sum_{j\in\mathcal{O}_i^{'}} \text{KL}(q_{ij}, q_{ij}^{'})\mathbb{E}_\nu\left[N_{ij}(T)\right] \leq \sum_{j\in\mathcal{O}_i^{'}} (\text{KL}(q_{ij}, 1/2) + \epsilon)\mathbb{E}_\nu\left[N_{ij}(T)\right].$$

From the Bretagnolle-Huber inequality, for any event $A \in \mathcal{F}_{M,T}$,

$$\begin{aligned}\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu'\pi}(A^c) &\geq \frac{1}{2}\exp\left(-\text{KL}\left(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi}\right)\right) \\ &\geq \frac{1}{2}\exp\left(-\sum_{j\in\mathcal{O}_i^{'}} (\text{KL}(q_{ij}, 1/2) + \epsilon)\mathbb{E}_\nu\left[N_{ij}(T)\right]\right).\end{aligned}$$

Choose $A = \{N_{ii}(T) > MT/2\}$, and let $\mathcal{R}_\mathcal{G} := \mathcal{R}_{\mathcal{G},T}(\pi, \nu), \mathcal{R}_\mathcal{G}^{'} := \mathcal{R}_{\mathcal{G},T}(\pi, \nu')$. By the Markov inequality,

$$\mathcal{R}_\mathcal{G} \geq \Delta_{1i}\mathbb{E}_\nu\left[N_{ii}(T)\right] \geq \frac{MT\Delta_{1i}}{2}\mathbb{P}_{\nu\pi}\left(N_{ii}(T) \geq \frac{MT}{2}\right) = \frac{MT\Delta_{1i}}{2}\mathbb{P}_{\nu\pi}(A).$$

Additionally, the regret for the challenging instance is

$$\begin{aligned}\mathcal{R}_\mathcal{G}^{'} = \frac{1}{2}\sum_{l\in[K]\setminus\{i\}}\sum_{j\in[K]\setminus\{i,l\}} &\frac{\Delta_{il}^{'} + \Delta_{ij}^{'}}{2}\mathbb{E}_{\nu'}N_{lj}(T) \\ + \sum_{j\in[K]} \Delta_{ij}^{'}\mathbb{E}_{\nu'}N_{jj}(T) + &\sum_{j\in[K]\setminus\{i\}} \frac{\Delta_{ij}^{'}}{2}\mathbb{E}_{\nu'}N_{ij}(T).\end{aligned}$$

By definition, since arm i is the CW for the challenging instance, the minimal gap with respect to it $\Delta_{min}^{'} =$

$\min_{j\in[K]\setminus\{i\}} \Delta^{'}_{ij}$ is positive, so

$$
\begin{aligned}
\mathcal{R}^{'}_{\mathcal{G}} &\geq \frac{1}{2}\sum_{l\in[K]\setminus\{i\}}\sum_{j\in[K]\setminus\{i,l\}} \frac{\Delta^{'}_{min}/2+\Delta^{'}_{min}/2}{2}\mathbb{E}_{\nu'}N_{lj}(T) \\
&+ \sum_{j\in[K]\setminus\{i\}}(\Delta^{'}_{min}/2)\mathbb{E}_{\nu'}N_{jj}(T) + \sum_{j\in[K]\setminus\{i\}}\frac{\Delta^{'}_{min}}{2}\mathbb{E}_{\nu'}N_{ij}(T) \\
&= \frac{\Delta^{'}_{min}}{2}\mathbb{E}_{\nu'}\left[\frac{1}{2}\sum_{l\in[K]\setminus\{i\}}\sum_{j\in[K]\setminus\{i,l\}}N_{lj}(T) + \sum_{j\in[K]\setminus\{i\}}N_{jj}(T) + \sum_{j\in[K]\setminus\{i\}}N_{ij}(T)\right] \\
&= \frac{\Delta^{'}_{min}}{2}\mathbb{E}_{\nu'}\left[MT - N_{ii}(T)\right] \geq \frac{\Delta^{'}_{min}}{2}(MT-\frac{MT}{2})\mathbb{P}_{\nu'\pi}\left(MT-N_{ii}(T)\geq MT-\frac{MT}{2}\right) \\
&= \frac{MT\Delta^{'}_{min}}{4}\mathbb{P}_{\nu'\pi}(A^c).
\end{aligned}
$$

Putting it all together,

$$
\begin{aligned}
\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}} &\geq \frac{MT\Delta_{1i}}{2}\mathbb{P}_{\nu\pi}(A) + \frac{MT\Delta^{'}_{min}}{4}\mathbb{P}_{\nu'\pi}(A^c) \\
&\geq \frac{MT}{4}\min\{2\Delta_{1i},\Delta^{'}_{min}\}\left(\mathbb{P}_{\nu\pi}(A)+\mathbb{P}_{\nu'\pi}(A^c)\right) \\
&\geq \frac{MT}{8}\min\{2\Delta_{1i},\Delta^{'}_{min}\}\exp\left(-\sum_{j\in\mathcal{O}^{'}_{i}}(\mathrm{KL}(q_{ij},1/2)+\epsilon)\mathbb{E}_{\nu}\left[N_{ij}(T)\right]\right).
\end{aligned}
$$

Rearranging the previous inequality,

$$
\sum_{j\in\mathcal{O}^{'}_{i}}(\mathrm{KL}(q_{ij},1/2)+\epsilon)\mathbb{E}_{\nu}\left[N_{ij}(T)\right] \geq \log\left(\frac{MT\min\{2\Delta_{1i},\Delta^{'}_{min}\}}{8\left(\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}}\right)}\right),
$$

and:

$$
\begin{aligned}
\liminf_{T\to\infty}\frac{\sum_{j\in\mathcal{O}^{'}_{i}}(\mathrm{KL}(q_{ij},1/2)+\epsilon)\mathbb{E}_{\nu}\left[N_{ij}(T)\right]}{\log T} &\geq \liminf_{T\to\infty}\frac{\log\left(\frac{MT\min\{2\Delta_{1i},\Delta^{'}_{min}\}}{8\left(\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}}\right)}\right)}{\log T} \\
&= \left(1-\limsup_{T\to\infty}\frac{\log\left(\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}}\right)}{\log T}\right) = 1.
\end{aligned}
$$

In the last transition we used that fact that $\pi\in\Pi_{\mathrm{cons}}(\mathcal{Q})$, so for every $p>0$ there exists some constant C_p such that $\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}}\leq C_pT^p$ for large enough T. This means that $\log\left(\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}}\right)\leq p\log T+\log C_p$, so

$$
\limsup_{T\to\infty}\frac{\log\left(\mathcal{R}_{\mathcal{G}}+\mathcal{R}^{'}_{\mathcal{G}}\right)}{\log T}\leq\limsup_{T\to\infty}\frac{p\log T+\log C_p}{\log T}=p.
$$

Since p is an arbitrary positive constant, by taking it to 0, the transition above holds. Similarly, ϵ is also an arbitrary positive constant, so by taking it to zero, all the arms for which $j\in\mathcal{O}^{'}_{i}\setminus\mathcal{O}_i$ disappear from the sum above, as $\mathrm{KL}(q_{ij},1/2)=0$. This concludes the proof. □

As the final part of this section, we provide a proof for Theorem 3.1.

Algorithm 3: Simple Leader Election

```
Input: Number of players $M$, communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
Initialize: $\mathrm{ID}(m)$ - unique ID in $[M]$ for $m \in \mathcal{V}$.
for $t = 1, \dots, D+1$ do
    for $m = 1, \dots, M$ do
        Send $\mathrm{ID}(m)$ to neighbors.
        Receive ID values originating from neighbors $\mathcal{N}(m)$.
        $\mathrm{ID}(m) \leftarrow \min\{\mathrm{ID}(m') \mid m' \in \mathcal{N}(m) \text{ or } m' = m\}$.
```

Proof. We use the definition of the regret and Lemma C.9,

$$\begin{aligned}
\mathbb{E}\mathcal{R}_{\mathcal{G},T}(\pi,\nu) &= \mathbb{E}_\nu\left[\frac{1}{2}\sum_{i\in[K]}\sum_{j\in[K]\setminus\{i\}}\frac{\Delta_{1i}+\Delta_{1j}}{2}N_{ij}(T) + \sum_{i\in[K]}\Delta_{1i}N_{ii}(T)\right] \\
&\geq \mathbb{E}_\nu\left[\sum_{i,j\in[K]:q_{ij}<1/2}\frac{\Delta_{1i}+\Delta_{1j}}{2}N_{ij}(T) + \sum_{i\in[K]}\Delta_{1i}N_{ii}(T)\right] \\
&\geq \mathbb{E}_\nu\left[\sum_{i\in[K]\setminus\{1\}}\sum_{j\in\mathcal{O}_i}\frac{\Delta_{1i}+\Delta_{1j}}{2}N_{ij}(T)\right] \\
&= \mathbb{E}_\nu\left[\sum_{i\in[K]\setminus\{1\}}\sum_{j\in\mathcal{O}_i}\frac{\Delta_{1i}+\Delta_{1j}}{2\mathrm{KL}(q_{ij},1/2)}\mathrm{KL}(q_{ij},1/2)N_{ij}(T)\right] \\
&\geq \sum_{i\in[K]\setminus\{1\}}\min_{j\in\mathcal{O}_i}\frac{\Delta_{1i}+\Delta_{1j}}{2\mathrm{KL}(q_{ij},1/2)}\mathbb{E}_\nu\left[\sum_{j\in\mathcal{O}_i}\mathrm{KL}(q_{ij},1/2)N_{ij}(T)\right].
\end{aligned}$$

□

D FOLLOW YOUR LEADER BLACK BOX ALGORITHM PROOFS

In this section, we prove and elaborate on the various claims from Section4.

D.1 A Simple Leader Election Algorithm

Since our focus in Theorem 4.2 is on the asymptotic regime, we narrow our attention to deterministic identified leader election algorithms, denoted as LEAlg, capable of completion within T_{LE} rounds, irrespective of the overall time horizon T. An illustrative example is presented in Algorithm 3, which also finds application in the multiplayer MAB setup discussed in Wang et al. (2020). Our assumption entails that each player is initialized with a unique deterministic ID in the range $[M]$. Subsequently, players share their IDs with neighbors and update their IDs based on the minimum value obtained from their own ID and ones that were received from neighbors. Consequently, by the end of the leader election phase, the player originally possessing the minimum ID assumes the role of the leader, while the remaining players become followers as their IDs undergo changes.

For this algorithm, leader election is completed within $T_{\mathrm{LE}} = D + 1$ rounds. Additional examples of applicable algorithms can be found in Casteigts et al. (2019).

D.2 Proof for the Claims in Section 4

We start with a proof for Theorem 4.2.

Proof. First, we discuss the random regret $\mathcal{R}(T)$. This can be decomposed into four parts:

- **Leader Election Phase:** For a deterministic T_{LE} number of rounds, each player draws some arm pair determined by the base algorithm. Taking into account the time required for the leader's CW candidate to reach all the players for the first time after the former's election, we have a contribution of $M(T_{\text{LE}}+D)\Delta_{1\max}$ to the regret bound.

- **Followers Exploitation Rounds:** For rounds where the followers draw the same arm as the CW candidate held by the leader, regret is incurred only if this arm is not the CW. This results in a regret bound contribution of $(M-1)\sum_{t=T_{\text{LE}}+1}^{T}\Delta_{1\max}\mathbb{1}\{\text{cw}(t)\neq 1\}$.

- **Communication Rounds:** These are rounds in which the arm drawn by followers differs from the CW candidate held by the leader due to communication delay, i.e., rounds where the updated arm information has not yet reached these followers. Given the maximal delay of D rounds, this contributes an additional factor of $MD\Delta_{1\max}$ to the regret bound, as in the worst case a given player draws at round T the leader's candidate from round $T-D$.

- **Leader's Regret:** The leader draws arms as indicated by the base algorithm, and by only utilizing its own samples, and the resulting contribution to the regret is denoted by $\mathcal{R}_{m_l}(T)$.

Using the decomposition above, we bound the random regret as,

$$\mathcal{R}(T)\leq M(T_{\text{LE}}+2D)\Delta_{1\max}+M\Delta_{1\max}\sum_{t=T_{\text{LE}}+1}^{T}\mathbb{1}\{\text{cw}(t)\neq 1\}+\mathcal{R}_{m_l}(T)\ .$$

To complete the proof for the average regret, let us denote $\mathbb{P}$ as the probability measure for the entire multiplayer system, and $\mathbb{P}'$ as the measure for a single-player system containing only the leader operating as indicated by the base algorithm. For any history $\mathcal{H}_t^{m_l}$ containing only the leader's draws and rewards, we show that it has the same probability under both measures. To this end, we use an induction principle. This holds trivially for $t=1$, and for $t>1$ we get that since the leader uses the base algorithm without information outside of $\mathcal{H}_t^{m_l}$, and since it is elected deterministically by the leader election algorithm, the following holds.

$$\begin{aligned}&P(\mathcal{H}_{t+1}^{m_l})=P((i_{m_l}(t+1),j_{m_l}(t+1)),r_{m_l}(t+1),\mathcal{H}_t^{m_l})\\&=P((i_{m_l}(t+1),j_{m_l}(t+1)),r_{m_l}(t+1)|\mathcal{H}_t^{m_l})P(\mathcal{H}_t^{m_l})\\&=P'((i_{m_l}(t+1),j_{m_l}(t+1)),r_{m_l}(t+1)|\mathcal{H}_t^{m_l})P'(\mathcal{H}_t^{m_l})=P'(\mathcal{H}_{t+1}^{m_l})\ ,\end{aligned}$$

where the third transition follows from the induction assumption and from the fact that the leader follows the base algorithm in both environments. Since neither $\mathcal{R}_{m_l}(T)$ nor $\text{cw}(t)$ contain information outside $\mathcal{H}_t^{m_l}$, Assumption 4.1 leads to,

$$\begin{aligned}\mathbb{E}\mathcal{R}(T)&\leq M(T_{\text{LE}}+2D)\Delta_{1\max}+M\Delta_{1\max}\mathbb{E}_{\mathbb{P}}\left[\sum_{t=T_{\text{LE}}+1}^{T}\mathbb{1}\{\text{cw}(t)\neq 1\}\right]+\mathbb{E}_{\mathbb{P}}\mathcal{R}_{m_l}(T)\\&=M(T_{\text{LE}}+2D)\Delta_{1\max}+M\Delta_{1\max}\mathbb{E}_{\mathbb{P}'}\left[\sum_{t=T_{\text{LE}}+1}^{T}\mathbb{1}\{\text{cw}(t)\neq 1\}\right]+\mathbb{E}_{\mathbb{P}'}\mathcal{R}_{m_l}(T)\\&\leq M(T_{\text{LE}}+2D)\Delta_{1\max}+M\Delta_{1\max}f(K,T,Q)+g(K,T,Q)\ .\end{aligned}$$

The decomposition above guarantees that communication, once the leader election phase is completed, is only initiated by the leader in rounds where $\text{cw}(t-1)\neq\text{cw}(t)$, and the expected number of such rounds is bounded by $2f$. This completes the proof for the expected regret bound. Additionally, this shows that the expected number of communication rounds is bounded by $T_{\text{LE}}+2f(K,T,Q)$. □

Next, to prove Corollary 4.3 we present the following lemma.

Lemma D.1. *For any $\epsilon>0$ and $\alpha>1$, the following holds.*

(a) *For RUCB (Zoghi et al., 2014) as the base algorithm, define the CW candidate* $\mathrm{cw}(t)$ *at each round as the hypothesized best arm* $\mathcal{B}(t)$ *if it is not empty. Otherwise, use a random arm. Then,*

$$f^{RUCB}(K,T,Q) = O\left(\frac{K^2 \log K}{\Delta^2_{1\min}}\right) \ ,$$
$$g^{RUCB}(K,T,Q) = \sum_{i=2}^{K} \frac{4\alpha \log T}{\Delta_{1i}} + O\left(\frac{K^2 \log K}{\Delta^2_{1\min}}\right) \ .$$

(b) *For RMED2FH (Komiyama et al., 2015) as the base algorithm, define the CW candidate* $\mathrm{cw}(t)$ *at each round to be the arm* $i^*(t) = \operatorname{argmin} I_i(t)$*, where* $I_i(t)$ *is the empirical divergence of arm* i*. Then,*

$$f^{RMED2FH}(K,T,Q) = O\left(K^2 \log\log T + K^{2+\epsilon}\right) \ ,$$
$$g^{RMED2FH}(K,T,Q) = \sum_{i=2}^{K} \min_{j\in\mathcal{O}_i} \frac{\Delta_{1i}+\Delta_{1j}}{2\mathrm{KL}(q_{ij},1/2)} \log T$$
$$+ O\left(K^2 \log\log T + K^{2+\epsilon} + \frac{K\log T}{\log\log T}\right) \ .$$

Proof. We begin with a proof for part 1, utilizing RUCB as a base algorithm. By definition, the bound g^{RUCB} is identical to the expected regret bound of RUCB. Following a proof scheme similar to that presented in Zoghi et al. (2014), this results in,

$$g^{RUCB}(K,T,Q) = \left[2\mathcal{D}\log 2\mathcal{D} + 2\left(K^2 \frac{4\alpha-1}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}} \frac{2\alpha-1}{\alpha-1}\right]\Delta_{1\max} + \sum_{i\in[K]/cw} \frac{4\alpha\log T}{\Delta_{1i}} \ .$$

Here $\mathcal{D} = \sum_{i<j} D_{ij}$, where $D_{1j} = \frac{4\alpha}{\Delta^2_{1j}}$ for $i = 1$ and $D_{ij} = \frac{4\alpha}{\min\{\Delta^2_{1i},\Delta^2_{1j}\}}$ for $i, j \neq 1$. For the bound f^{RUCB}, with probability larger than $1-\delta$ it holds that the hypothesized CW $\mathcal{B}(t)$ contains the CW for all $t \geq \hat{T}_\delta$. Therefore,

$$\sum_{t=1}^{T} \mathbb{1}\{\mathrm{cw}(t) \neq 1\} \leq \hat{T}_\delta \leq 2\mathcal{D}\log 2\mathcal{D} + 2C(\delta) \ ,$$

where we use some notations from the proof scheme in Zoghi et al. (2014). By using a similar integration technique to the one utilized in Appendix E to establish a bound for the expected regret given a high probability regret bound, we obtain that,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathrm{cw}(t) \neq 1\}\right] \leq 2\mathcal{D}\log 2\mathcal{D} + 2\left(K^2 \frac{4\alpha-1}{2\alpha-1}\right)^{\frac{1}{2\alpha-1}} \frac{2\alpha-1}{\alpha-1} \ .$$

The proof for part one is concluded by observing that $\mathcal{D} = O(K^2 \log K/\Delta^2_{1\min})$.

For the second part, the bound $g^{RMED2FH}$ follows from the expected regret bound provided in Komiyama et al. (2015). For the bound $f^{RMED2FH}$, define the event

$$\mathcal{U}(t) := \cap_{i\in[K]\setminus 1}\{\hat{q}_{1,i}(t) > 1/2\} \ .$$

By definition, when $\mathcal{U}(t)$ is true we have $I_i(t) > 0$ for any arm $i \neq 1$, making $i^*(t) = 1$ unique with $I_{i^*(t)} = 0$. Therefore,

$$\sum_{t=1}^{T} \mathbb{1}\{\mathrm{cw}(t) \neq 1\} \leq \sum_{t=1}^{T} \mathbb{1}\{\mathcal{U}^c(t)\} \ .$$

From Lemma 5 in Komiyama et al. (2015),

$$\mathbb{E}\left[\sum_{t=T_{init}+1}^{T} \mathbb{1}\{\mathcal{U}^c(t)\}\right] = O\left(K^{2+\epsilon}\right) \ ,$$

for any $\epsilon > 0$. Finally, by substituting the value of T_{init} for RMED2FH, it holds that,

$$f^{RMED2FH}(K,T,Q) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathrm{cw}(t) \neq 1\}\right] = O\left(K^2 \log\log T + K^{2+\epsilon}\right) \ .$$

□

Corollary 4.3 in the main paper follows directly from the combination of the previous Lemma and Theorem 4.2.

E MESSAGE-PASSING RUCB PROOFS

In this section, we elaborate on the proof scheme for the theorem and lemmas presented in Section 5 and provide a slightly more general formulation that allows for more flexibility in the choice of constants.

For clarity, it is important to note that the UCB indices calculated by each player m for all $i \neq j$ are expressed as follows.

$$u_{ij}^m(t) := \frac{\tilde{w}_{ij}^m(t-1)}{\tilde{N}_{ij}^m(t-1)} + \sqrt{\frac{\alpha \log t}{\tilde{N}_{ij}^m(t-1)}} \ .$$

Using this notation, at the beginning of round t, players employ $\tilde{w}_{ij}^m(t-1), \tilde{N}_{ij}^m(t-1)$ to update the UCB terms. The values of $\tilde{w}_{ij}^m(t), \tilde{N}_{ij}^m(t)$ are then updated at the end of the round after the communication phase is concluded.

In the proof scheme presented in this section, we utilize a different probability space compared to the one defined in SectionC. We introduce a collection of independent Bernoulli random variables $(q_{ij}^m(t))$ for $t \in [T], i,j \in [K], m \in [M]$, satisfying $\mathbb{E}q_{ij}^m(t) = q_{ij}$. Caratheodory's extension theorem allows us to define this collection of random variables on the probability space $(\Omega, \mathcal{F}) = \left(\mathbb{R}^{TMK^2}, \mathcal{B}\left(\mathbb{R}^{TMK^2}\right)\right)$. Under this model, the reward $r_m(t)$ collected by player m at round t is represented as $r_m(t) = q_{i_m(t)j_m(t)}^m(t)$. By defining the filtration $\mathcal{F}_t = \sigma\left(q_{ij}^m(\tau)\right)_{\tau \le t}$, we ensure that $i_m(t), j_m(t) \in \mathcal{F}_{t-1}$ and $r_m(t) \in \mathcal{F}_t$. Subsequently, we define $r_m^{ij}(t)$ as the reward obtained by player m, given that the pair (i,j) was drawn at round t. Otherwise, it is defined to be 0.

$$r_{ij}^m(t) = q_{ij}^m(t)\mathbb{1}\{(i_m(t), j_m(t)) = (i,j)\} + (1 - q_{ji}^m(t))\mathbb{1}\{(i_m(t), j_m(t)) = (j,i)\} \ ,$$

In addition, define the LCB terms for all i,j,m as,

$$l_{ij}^m(t) = 1 - u_{ij}^m(t) \ .$$

We begin by presenting a concentration bound in the following lemma.

Lemma E.1. *The following concentration bounds hold for all $i, j, m, t > 0, \eta > 1$ and $\delta > 0$.*

$$P\left(q_{ij} < l_{ij}^m(t)\right) < \frac{\log\left(d_m(\mathcal{G}_\gamma) + 1\right) t}{t^{2\alpha\eta^{-1/2}} \log \eta}$$

$$P\left(q_{ij} > u_{ij}^m(t)\right) < \frac{\log\left(d_m(\mathcal{G}_\gamma) + 1\right) t}{t^{2\alpha\eta^{-1/2}} \log \eta} \ ,$$

where $\mathcal{G}_\gamma$ stands for the γ-power of the graph $\mathcal{G}$, and $d_m(\cdot)$ is the degree of node m on the graph.

Proof. For $i = j$, this holds by definition. Going forward, we concentrate on the case $i \neq j$. For simplicity, let us denote

$$\mathbb{1}^{m,m'}(t,\tau) := \mathbb{1}\{d(m,m') \le \min(\gamma, t - \tau + 1)\} \ ,$$

and

$$\chi_{ij}^{m,m'}(t,\tau) := \mathbb{1}\{(i_{m'}(\tau), j_{m'}(\tau)) = (i,j)\}\mathbb{1}^{m,m'}(t,\tau) \ .$$

This indicator determines whether the arm pair (i,j) was drawn by player m' at round τ and whether it is included in the statistics of player m at time t. The latter condition is satisfied only if the distance between the nodes on the graph is smaller than γ and if the time difference between t and the drawing time τ is greater than the time it takes to traverse the graph $(d(m,m')-1)$. Next, for $\tau \leq t$ define,

$$\tilde{r}_{ij}^{m,m'}(t,\tau) := q_{ij}^{m'}(\tau)\chi_{ij}^{m,m'}(t,\tau) + (1-q_{ji}^{m'}(\tau))\chi_{ji}^{m,m'}(t,\tau) \ ,$$

and,

$$Y_{ij}^{m}(t,\tau) := \sum_{m'=1}^{M}\left(\tilde{r}_{ij}^{m,m'}(t,\tau) - \mathbb{E}\left[\tilde{r}_{ij}^{m,m'}(t,\tau)|\mathcal{F}_{\tau-1}\right]\right) \ .$$

For any player m and arm pair (i,j), $Y_{ij}^{m}(t,\tau)$ represents the contribution at round t of conditional centralized rewards drawn at round τ from all players from which the message can be received at the current round. Since the first indicator in the definition of $\chi_{ij}^{m,m'}(t,\tau)$ only depends on draws at round τ and the second indicator is deterministic, it holds that $\chi_{ij}^{m,m'}(t,\tau) \in \mathcal{F}_{\tau-1}$. Therefore, we can decompose $Y_{ij}^{m}(t,\tau)$ as follows.

$$Y_{ij}^{m}(t,\tau) = \sum_{m'=1}^{M}\left((q_{ij}^{m'}(\tau)-q_{ij})\chi_{ij}^{m,m'}(t,\tau) + (1-q_{ji}^{m'}(\tau)-q_{ij})\chi_{ji}^{m,m'}(t,\tau)\right) \ .$$

By taking an expected value, it holds that for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda Y_{ij}^{m}(t,\tau)\right)|\mathcal{F}_{\tau-1}\right] \\
&= \mathbb{E}\left[\exp\left(\lambda\sum_{m'=1}^{M}\left((q_{ij}^{m'}(\tau)-q_{ij})\chi_{ij}^{m,m'}(t,\tau) + (1-q_{ji}^{m'}(\tau)-q_{ij})\chi_{ji}^{m,m'}(t,\tau)\right)\right)\middle|\mathcal{F}_{\tau-1}\right] \\
&= \prod_{m'=1}^{M}\mathbb{E}\left[\exp\left(\lambda\left((q_{ij}^{m'}(\tau)-q_{ij})\chi_{ij}^{m,m'}(t,\tau) + (1-q_{ji}^{m'}(\tau)-q_{ij})\chi_{ji}^{m,m'}(t,\tau)\right)\right)\middle|\mathcal{F}_{\tau-1}\right] \\
&= \prod_{m'=1}^{M}\mathbb{E}\left[\exp\left(\lambda(q_{ij}^{m'}(\tau)-q_{ij})\chi_{ij}^{m,m'}(t,\tau)\right)|\mathcal{F}_{\tau-1}\right] \\
&\times \prod_{m'=1}^{M}\mathbb{E}\left[\exp\left(\lambda(1-q_{ji}^{m'}(\tau)-q_{ij})\chi_{ji}^{m,m'}(t,\tau)\right)|\mathcal{F}_{\tau-1}\right] \\
&\leq \prod_{m'=1}^{M}\exp\left(\frac{\lambda^2}{8}\chi_{ij}^{m,m'}(t,\tau)\right)\exp\left(\frac{\lambda^2}{8}\chi_{ji}^{m,m'}(t,\tau)\right) \\
&= \prod_{m'=1}^{M}\exp\left(\frac{\lambda^2}{8}\left(\chi_{ij}^{m,m'}(t,\tau)+\chi_{ji}^{m,m'}(t,\tau)\right)\right) \\
&= \exp\left(\frac{\lambda^2}{8}\sum_{m'=1}^{M}\zeta_{ij}^{m,m'}(t,\tau)\right) \ ,
\end{aligned}$$

where

$$\zeta_{ij}^{m,m'}(t,\tau) := \zeta_{ij}^{m'}(\tau)\mathbb{1}^{m,m'}(t,\tau) \ ,$$

and $\zeta_{ij}^{m}(t)$ is an indicator specifying whether the pair (i,j) was drawn by player m at round t in any order,

$$\zeta_{ij}^{m}(t) := \mathbb{1}\{(i_m(t),j_m(t)) \in \{(i,j),(j,i)\}\} \ .$$

In the second transition above, we leverage the independence of $q_{ij}^{m}(t)$ across different players and arms, and the fact that all other terms are determined by conditioning on $\mathcal{F}_{\tau-1}$. The third transition follows a similar logic. For the fourth transition, we use the fact that both $q_{ij}^{m}(t)-q_{ij}$ and $1-q_{ji}^{m}(t)-q_{ij}$ are centered Bernoulli random variables, and as such, they are also $1/2-$sub-Gaussian. The inequality above implies that,

$$\mathbb{E}\left[\exp\left(\lambda Y_{ij}^{m}(t,\tau) - \frac{\lambda^2}{8}\sum_{m'=1}^{M}\zeta_{ij}^{m,m'}(t,\tau)\right)|\mathcal{F}_{\tau-1}\right] \leq 1 \ .$$

Note that the visitation counter utilized by players can be decomposed as the following sum of contributions from all players and prior rounds.

$$\tilde{N}_{ij}^{m}(t) = \sum_{\tau=1}^{t} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,\tau) \ .$$

Next, define

$$Z_{ij}^{m}(t) := \sum_{\tau=1}^{t} Y_{ij}^{m}(t,\tau) \ .$$

It holds that,

$$\begin{aligned} Y_{ij}^{m}(t,\tau) &= \sum_{m'=1}^{M} \left(r_{ij}^{m'}(\tau) \mathbb{1}^{m,m'}(t,\tau) - q_{ij} \zeta_{ij}^{m'}(\tau) \mathbb{1}^{m,m'}(t,\tau) \right) \\ &= \sum_{m'=1}^{M} r_{ij}^{m'}(\tau) \mathbb{1}^{m,m'}(t,\tau) - q_{ij} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,\tau) \ , \end{aligned}$$

meaning that $Z_{ij}^{m}(t)$ is a random variable representing the centered number of wins,

$$Z_{ij}^{m}(t) = \tilde{w}_{ij}^{m}(t) - q_{ij} \tilde{N}_{ij}^{m}(t) \ .$$

Employing the previous inequality, it holds that,

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\lambda Z_{ij}^{m}(t) - \frac{\lambda^2}{8} \tilde{N}_{ij}^{m}(t) \right) \middle| \mathcal{F}_{t-1} \right] \\ &= \mathbb{E}\left[\exp\left(\lambda Y_{ij}^{m}(t,t) - \frac{\lambda^2}{8} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,t) \right.\right. \\ &\qquad \left.\left. + \lambda \sum_{\tau=1}^{t-1} Y_{ij}^{m}(t,\tau) - \frac{\lambda^2}{8} \sum_{\tau=1}^{t-1} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,\tau) \right) \middle| \mathcal{F}_{t-1} \right] \\ &= \mathbb{E}\left[\exp\left(\lambda Y_{ij}^{m}(t,t) - \frac{\lambda^2}{8} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,t) \right) \middle| \mathcal{F}_{t-1} \right] \\ &\qquad \times \exp\left(\lambda \sum_{\tau=1}^{t-1} Y_{ij}^{m}(t,\tau) - \frac{\lambda^2}{8} \sum_{\tau=1}^{t-1} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,\tau) \right) \\ &\le \exp\left(\lambda \sum_{\tau=1}^{t-1} Y_{ij}^{m}(t,\tau) - \frac{\lambda^2}{8} \sum_{\tau=1}^{t-1} \sum_{m'=1}^{M} \zeta_{ij}^{m,m'}(t,\tau) \right) \ , \end{aligned}$$

where the second transition follows the fact that samples collected before time t are determined by $\mathcal{F}_{t-1}$. By extending this inequality further in time and applying the tower rule, we obtain the following.

$$\mathbb{E}\left[\exp\left(\lambda Z_{ij}^{m}(t) - \frac{\lambda^2}{8} \tilde{N}_{ij}^{m}(t) \right) \right] \le 1 \ .$$

Next, from the Markov inequality it holds that for any $\theta, \kappa, \lambda \in \mathbb{R}$,

$$e^{-2\kappa\theta} \ge P\left(\exp\left(\lambda Z_{ij}^{m}(t) - \frac{\lambda^2}{8} \tilde{N}_{ij}^{m}(t) \right) \ge e^{2\kappa\theta} \right) = P\left(\lambda Z_{ij}^{m}(t) - \frac{\lambda^2}{8} \tilde{N}_{ij}^{m}(t) \ge 2\kappa\theta \right) \ .$$

While this is valid for any $\lambda \in \mathbb{R}$, dividing the subsequent steps into segments for $\lambda > 0$ and $\lambda < 0$ will facilitate the proofs of the first and second concentration bounds, respectively. Given the similarity between the two proofs, we concentrate on the case $\lambda > 0$ in the following. It is evident that,

$$P\left(\frac{Z_{ij}^{m}(t)}{\sqrt{\tilde{N}_{ij}^{m}(t)}} \ge \frac{2\kappa\theta}{\lambda\sqrt{\tilde{N}_{ij}^{m}(t)}} + \frac{\lambda}{8}\sqrt{\tilde{N}_{ij}^{m}(t)} \right) \le e^{-2\kappa\theta} \ .$$

Next, we note that for any pair of arms (i,j), player m can receive no more than $d_m(\mathcal{G}_\gamma)$ samples at each round from other players. Hence, $1 \leq \tilde{N}_{ij}^m(t) \leq (1 + d_m(\mathcal{G}_\gamma))\, t$. Utilizing this fact, for any $\eta > 1$, it holds that $1 \leq \tilde{N}_{ij}^m(t) \leq \eta^{D_t}$ for

$$D_t := \frac{\log\left(1 + d_m(\mathcal{G}_\gamma)\right) t}{\log \eta} \ .$$

This enables us to address the randomness in $Z_{ij}^m(t), \tilde{N}_{ij}^m(t)$ by employing a peeling argument and partitioning the potential values of $\tilde{N}_{ij}^m(t)$ into intervals $[\eta^{l-1}, \eta^l]$ for $1 \leq l \leq D_t$. For this purpose, we introduce

$$\lambda_l := 4\sqrt{\frac{\kappa\theta}{\eta^{l-0.5}}} \ , \kappa = \eta^{-\frac{1}{2}} \ .$$

For $\tilde{N}_{ij}^m(t) \in [\eta^{l-1}, \eta^l]$ it holds that,

$$\frac{2\kappa\theta}{\lambda_l\sqrt{\tilde{N}_{ij}^m(t)}} + \frac{\lambda_l}{8}\sqrt{\tilde{N}_{ij}^m(t)} = \frac{\sqrt{\kappa\theta}}{2}\left(\sqrt{\frac{\eta^{l-0.5}}{\tilde{N}_{ij}^m(t)}} + \sqrt{\frac{\tilde{N}_{ij}^m(t)}{\eta^{l-0.5}}}\right) \leq \sqrt{\theta} \ ,$$

and by summing over the different intervals,

$$\begin{aligned}
P\left(\frac{Z_{ij}^m(t)}{\sqrt{\tilde{N}_{ij}^m(t)}} \geq \sqrt{\theta}\right) &\leq \sum_{l=1}^{D_t} P\left(\frac{Z_{ij}^m(t)}{\sqrt{\tilde{N}_{ij}^m(t)}} \geq \sqrt{\theta}, \tilde{N}_{ij}^m(t) \in [\eta^{l-1}, \eta^l]\right) \\
&\leq \sum_{l=1}^{D_t} P\left(\frac{Z_{ij}^m(t)}{\sqrt{\tilde{N}_{ij}^m(t)}} \geq \frac{2\kappa\theta}{\lambda_l\sqrt{\tilde{N}_{ij}^m(t)}} + \frac{\lambda_l}{8}\sqrt{\tilde{N}_{ij}^m(t)},\ \tilde{N}_{ij}^m(t) \in [\eta^{l-1}, \eta^l]\right) \\
&\leq \sum_{l=1}^{D_t} P\left(\frac{Z_{ij}^m(t)}{\sqrt{\tilde{N}_{ij}^m(t)}} \geq \frac{2\kappa\theta}{\lambda_l\sqrt{\tilde{N}_{ij}^m(t)}} + \frac{\lambda_l}{8}\sqrt{\tilde{N}_{ij}^m(t)}\right) \\
&\leq \sum_{l=1}^{D_t} e^{-2\kappa\theta} = \frac{\log\left(1 + d_m(\mathcal{G}_\gamma)\right) t}{\log \eta} e^{-2\kappa\theta} \ .
\end{aligned}$$

To conclude, substitute $\theta = \alpha \log t$,

$$\begin{aligned}
P\left(q_{ij} \leq l_{ij}^m(t+1)\right) &= P\left(\frac{\tilde{w}_{ij}^m(t)}{\tilde{N}_{ij}^m(t)} - q_{ij} \geq \sqrt{\frac{\alpha \log t}{\tilde{N}_{ij}^m(t)}}\right) \\
&= P\left(\frac{Z_{ij}^m(t)}{\tilde{N}_{ij}^m(t)} \geq \sqrt{\frac{\alpha \log t}{\tilde{N}_{ij}^m(t)}}\right) \leq \frac{\left(1 + d_m(\mathcal{G}_\gamma)\right)(t+1)}{\log \eta} \frac{1}{t^{2\alpha\eta^{-1/2}}} \ .
\end{aligned}$$

The concentration bound in the theorem results from replacing $t+1$ with t. □

The next lemma demonstrates that with high probability, $q_{ij} \leq u_{ij}^m(t)$ for all players, arm pairs, and rounds that are sufficiently large.

Lemma E.2. *For any $\epsilon > 0, \delta > 0$, and α such that $\alpha > 0.5\sqrt{\eta}(1+\epsilon)$, define*

$$T_\epsilon = \begin{cases} \left(\frac{-W_{-1}(-\epsilon)}{\epsilon}\right)^{\frac{1}{\epsilon}} & \text{if } \epsilon \in (0, e^{-1}) \\ 0 & \text{if } \epsilon \geq e^{-1} \ , \end{cases}$$

where $W_{-1}(\cdot)$ is the lower part of the Lambert function. Additionally, define

$$C_\epsilon(\delta) := \max\left(e^{\frac{1}{\epsilon}}, T_\epsilon, \left(\frac{MK^2\left(3 + 2\log(d_{\max}^\gamma + 1)\right)}{\min\{1, 2\alpha\eta^{-1/2} - 1\}\delta \log \eta}\right)^{\frac{1}{2\alpha\eta^{-1/2} - 1 - \epsilon}}\right) \ ,$$

where $d^\gamma_{\max} := \max_m \left(d_m(\mathcal{G}_\gamma)\right)$ *. UCB terms* $u^m_{ij}(t)$ *satisfy,*

$$P\left(\forall t \geq C_\epsilon(\delta), i, j, m:\ q_{ij} \leq u^m_{ij}(t)\right) > 1-\delta\ .$$

Proof. Utilizing Lemma E.1, it holds that,

$$\begin{aligned}
&P\left(\exists t \geq C_\epsilon(\delta), i, j, m:\ q_{ij} > u^m_{ij}(t)\right)\\
&\leq MK^2 \sum_{t=C_\epsilon(\delta)}^{\infty} \frac{\log\left(d^\gamma_{\max}+1\right)t}{t^{2\alpha\eta^{-1/2}}\log\eta}\\
&\leq \frac{MK^2}{\log\eta} \sum_{t=C_\epsilon(\delta)}^{\infty} \left(\frac{\log t}{t^{2\alpha\eta^{-1/2}}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{t^{2\alpha\eta^{-1/2}}}\right)\\
&\leq \frac{MK^2}{\log\eta}\left[\frac{\log C_\epsilon(\delta)}{C_\epsilon(\delta)^{2\alpha\eta^{-1/2}}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{C_\epsilon(\delta)^{2\alpha\eta^{-1/2}}} + \int_{C_\epsilon(\delta)}^{\infty}\left(\frac{\log t}{t^{2\alpha\eta^{-1/2}}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{t^{2\alpha\eta^{-1/2}}}\right)dt\right]\ ,
\end{aligned}$$

where we took advantage of the fact that for $t > C_\epsilon(\delta) > e$, the functions $\log t/t^{2\alpha\eta^{-1/2}}$ and $1/t^{2\alpha\eta^{-1/2}}$ are decreasing. Calculating the integrals above,

$$\begin{aligned}
&P\left(\exists t \geq C_\epsilon(\delta), i, j, m:\ q_{ij} > u^m_{ij}(t)\right)\\
&\leq \frac{MK^2}{\log\eta}\left[\frac{\log C_\epsilon(\delta)}{C_\epsilon(\delta)^{2\alpha\eta^{-1/2}}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{C_\epsilon(\delta)^{2\alpha\eta^{-1/2}}} + \right.\\
&\qquad \left. + \frac{(2\alpha\eta^{-1/2}-1)\log C_\epsilon(\delta)+1}{(2\alpha\eta^{-1/2}-1)^2 C_\epsilon(\delta)^{2\alpha\eta^{-1/2}-1}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{(2\alpha\eta^{-1/2}-1)C_\epsilon(\delta)^{2\alpha\eta^{-1/2}-1}}\right]\ .
\end{aligned}$$

For clarity, we denote $\overline{T} := C_\epsilon(\delta)$ in the remainder of this proof. To evaluate the right-hand side above, we are interested in the approximation $\log\overline{T} \leq \overline{T}^\epsilon$ for some $\epsilon > 0$. For $\epsilon \geq e^{-1}$, this holds for all $\overline{T} > 0$, and for $\epsilon \in (0, e^{-1})$ this is true for $\overline{T} \geq \left(-W_{-1}(-\epsilon)/\epsilon\right)^{1/\epsilon}$, where $W_{-1}(\cdot)$ stands for the lower part of the Lambert function. Therefore, when this inequality holds,

$$\begin{aligned}
&P\left(\exists t \geq C_\epsilon(\delta), i, j, m:\ q_{ij} > u^m_{ij}(t)\right)\\
&\leq \frac{MK^2}{\log\eta}\left[\frac{\log\overline{T}}{\overline{T}^{2\alpha\eta^{-1/2}}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{\overline{T}^{2\alpha\eta^{-1/2}}}\right.\\
&\qquad \left. + \frac{2\log\overline{T}}{(2\alpha\eta^{-1/2}-1)\overline{T}^{2\alpha\eta^{-1/2}-1}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{(2\alpha\eta^{-1/2}-1)\overline{T}^{2\alpha\eta^{-1/2}-1}}\right]\\
&\leq \frac{MK^2}{\log\eta}\left[\frac{1}{\overline{T}^{2\alpha\eta^{-1/2}-1-\epsilon}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{\overline{T}^{2\alpha\eta^{-1/2}-1-\epsilon}}\right.\\
&\qquad \left. + \frac{2}{(2\alpha\eta^{-1/2}-1)\overline{T}^{2\alpha\eta^{-1/2}-1-\epsilon}} + \frac{\log\left(d^\gamma_{\max}+1\right)}{(2\alpha\eta^{-1/2}-1)\overline{T}^{2\alpha\eta^{-1/2}-1-\epsilon}}\right]\\
&\leq \frac{MK^2\left(3+2\log\left(d^\gamma_{\max}+1\right)\right)}{\min\{1, 2\alpha\eta^{-1/2}-1\}\log\eta}\frac{1}{\overline{T}^{2\alpha\eta^{-1/2}-1-\epsilon}}\ ,
\end{aligned}$$

where we also used the fact that $\overline{T} > e^{1/\epsilon}$. By requiring the right-hand side to be smaller than δ, we have,

$$C_\epsilon(\delta) \geq \left(\frac{MK^2\left(3+2\log\left(d^\gamma_{\max}+1\right)\right)}{\min\{1, 2\alpha\eta^{-1/2}-1\}\delta\log\eta}\right)^{\frac{1}{2\alpha\eta^{-1/2}-1-\epsilon}}\ .$$

The proof concludes by combining the requirements on $C_\epsilon(\delta)$ mentioned above. □

Next, we show that with high probability, players in a clique cannot visit arm pairs different from $(1,1)$ for too many rounds. For brevity, we define,

$$D_{ij} := \begin{cases} 4\alpha/\overline{\Delta}^2_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j\ . \end{cases}$$

Lemma E.3. *For any $\epsilon > 0, \delta > 0, \gamma' \leq \gamma$, $T_0 \geq C_\epsilon(\delta)$ and α such that $\alpha > 0.5\sqrt{\eta}(1+\epsilon)$, let $\mathcal{C}(\mathcal{G}_{\gamma'})$ be the set of all cliques in $\mathcal{G}_{\gamma'}$. The number of visitations satisfies,*

$$P\left(\exists t > T_0, (i,j) \neq (1,1), \gamma' \leq \gamma, C \in \mathcal{C}(\mathcal{G}_{\gamma'}) : \sum_{m \in C} N_{ij}^{m,T_0}(t) > D_{ij} \log t + (\gamma' + 2)|C|\right) < \delta \ .$$

Proof. For the remainder of this proof, we assume that the good event $\{\forall t \geq C_\epsilon(\delta), i, j, m : \ q_{ij} \leq u_{ij}^m(t)\}$ holds, which occurs with probability $1 - \delta$ according to Lemma E.2. First, for draws of identical arms $(i,i) \neq (1,1)$ during rounds $t > C_\epsilon(\delta)$, the RUCB decision rule stipulates that player m can only draw arm i as both the first and the second arm if $u_{ji}^m(t) < 0.5$ for all $j \neq i$, since it holds that $u_{ii}^m(t) = 0.5$. However, given the good event, $u_{1i}^m(t) \geq q_{1i} > 0.5$, which is a contradiction. Hence, pairs of identical arms, apart from $(1,1)$, cannot be drawn under this event. Consequently, for any $C \in \mathcal{C}(\mathcal{G}_{\gamma'})$ it holds that $\sum_{m \in C} N_{ii}^{m,T_0}(t) = 0 < D_{ii} \log t + (\gamma' + 1)|C|$.

For pairs of non-identical arms $i \neq j$ with some given $C \in \mathcal{C}(\mathcal{G}_{\gamma'})$, let us assume that these arms have been compared for a sufficiently large number of rounds between T_0 and t, such that $\sum_{m \in C} N_{ij}^{m,T_0}(t-1) > D_{ij} \log t + (\gamma' + 1)|C|$. Note that this also dictates $t \geq \gamma' + 1 + T_0$. Let s denote the last round in which this pair was drawn by any player in C, such that $s \geq T_0 + \gamma' + 1$ [1]. Next, Consider several cases for the draw at round s, and notice that Algorithm 2 dictates that the first arm for each player m is always chosen from the winners set $\mathcal{C}^m$, which necessarily contains the CW and is thus non-empty.

- For $i, j \neq 1$, assume that $i_m(s) = i, j_m(s) = j$. It holds that $u_{ij}^m(s) \geq 0.5$, and $l_{ij}^m(s) := 1 - u_{ji}^m(s) \leq 1 - u_{1i}^m(s) \leq 1 - q_{1i}(s)$. Therefore, $u_{ij}^m(s) - l_{ij}^m(s) \geq \Delta_{1i}$.
- For $i, j \neq 1$ and $i_m(s) = j, j_m(s) = i$, a similar derivation reveals that $u_{ji}^m(s) - l_{ji}^m(s) \geq \Delta_{1j}$.
- For $i_m(s) = 1, j_m(s) = i$ to be drawn, it holds that $u_{1i}^m(s) \geq q_{1i}$ and $l_{1i}^m(s) = 1 - u_{i1}^m(s) \leq 0.5$. Hence, $u_{1i}^m(s) - l_{1i}^m(s) \geq \Delta_{1i}$.
- For $i_m(s) = i, j_m(s) = 1$ to be drawn, it must happen that $u_{i1}^m(s) \geq 0.5$ and $l_{i1}^m(s) = 1 - u_{1i}^m(s) \leq 1 - q_{1i}$. Therefore, it again holds that $u_{i1}^m(s) - l_{i1}^m(s) \geq \Delta_{1i}$.

To conclude, for $i \neq j$, a draw of the pair (i,j) at round s leads to $u_{ij}^m(s) - l_{ij}^m(s) \geq \min\{\Delta_{1i}, \Delta_{1j}\}$. Similarly, when one of the arms in the pair is the CW, it holds that $u_{i1}^m(s) - l_{i1}^m(s) \geq \Delta_{1i}$. On the other hand, by definition,

$$u_{ij}^m(s) - l_{ij}^m(s) = u_{ij}^m(s) + u_{ji}^m(s) - 1 = 2\sqrt{\frac{\alpha \log s}{\tilde{N}_{ij}^m(s-1)}} \ .$$

Utilizing the assumption about the number of visits, it holds that,

$$\begin{aligned}\tilde{N}_{ij}^m(s-1) &\geq \tilde{N}_{ij}^{m,T_0}(s-1) \\ &\geq \sum_{m \in C} N_{ij}^{m,T_0}(s - \gamma' - 1) \geq \sum_{m \in C} N_{ij}^{m,T_0}(s-1) - \gamma'|C| \\ &> D_{ij} \log t + (\gamma' + 1)|C| - |C| - \gamma'|C| \\ &= D_{ij} \log t \ ,\end{aligned}$$

where in the second transition, we leverage the fact that after γ' rounds, player m will have received all information available in C in the current round. The third transition takes advantage of the restriction that each player in the clique can only draw at most one pair at a round. Finally, the fourth transition relies on our assumption regarding the number of visits, along with the knowledge that s was the last round the arm pair was drawn by any player until round t. Plugging this above,

$$u_{ij}^m(s) - l_{ij}^m(s) < 2\sqrt{\frac{\alpha}{D_{ij}}} = \overline{\Delta}_{ij} \ .$$

By the definition of $\overline{\Delta}_{ij}$, this is a contradiction. Therefore, under the good event $\sum_{m \in C} N_{ij}^{m,T_0}(t-1) \leq D_{ij} \log t + (\gamma' + 1)|C|$ for all $t > T_0$, $(i,j) \neq (1,1), C \in \mathcal{C}(\mathcal{G}_{\gamma'})$. Hence, $\sum_{m \in C} N_{ij}^{m,T_0}(t) \leq D_{ij} \log t + (\gamma' + 2)|C|$, with probability $1 - \delta$. □

[1] During round s, several players may have drawn this pair.

Next, we demonstrate how the proof to Lemma 5.1 in the main paper follows from the previous lemma.

Proof. Take $\eta = 1.3$ and $\epsilon = e^{-1}$. For these values, it holds that $T_\epsilon = 0$, and by demanding $\alpha > 1.2$ we have $2\alpha\eta^{-1/2} - 1 > 1$, which makes the requirement $\geq e^{1/\epsilon}$ in the proof to Lemma E.2 irrelevant. By combining all of the above, we have that,

$$C(\delta) := \left(\frac{4MK^2\left(3 + 2\log\left(d_{\max}^{\gamma} + 1\right)\right)}{\delta}\right)^{\frac{1}{1.7\alpha - 1.4}}$$

is sufficient for the concentration bound in Lemma E.2 to hold. By utilizing these in Lemma E.3, we derive the version found in the main paper. □

Lemma E.4. *For any $\epsilon > 0, \delta > 0, a > 0$, $\gamma' \leq \gamma$ and α such that $\alpha > 0.5\sqrt{\eta}(1+\epsilon)$, let $\hat{T}_\epsilon(\delta)$ be the smallest round satisfying*

$$\hat{T}_\epsilon(\delta) > h(K, \gamma', T_0) + a\log\hat{T}_\epsilon(\delta)\ ,$$

where $h(\cdot)$ is independent of $\hat{T}_\epsilon(\delta)$. Then, $\hat{T}_\epsilon(\delta)$ exists, and it holds that,

$$\hat{T}_\epsilon(\delta) \leq 2h(K, \gamma', T_0) + 2a\log(2a)\ .$$

Proof. To prove the lemma, it is sufficient to find a value of τ such that $\tau > h(K, \gamma', T_0) + a\log\tau$. By the definition of $\hat{T}_\epsilon(\delta)$, this means that $\hat{T}_\epsilon(\delta) < \tau$. We will demonstrate that $\tau = 2h(K, \gamma', T_0) + 2a\log(2a)$ satisfies this condition. Using basic algebra,

$$\begin{aligned}
&h(K, \gamma', T_0) + a\log\tau \\
&= h(K, \gamma', T_0) + a\log\left(2h(K, \gamma', T_0) + 2a\log(2a)\right) \\
&= h(K, \gamma', T_0) + a\log\left(2a\log(2a)\right) + a\log\left(1 + \frac{2h(K, \gamma', T_0)}{2a\log(2a)}\right) \\
&\leq h(K, \gamma', T_0) + a\log\left(2a\log(2a)\right) + a\cdot\frac{2h(K, \gamma', T_0)}{2a\log(2a)} \\
&\leq h(K, \gamma', T_0) + a\log\left((2a)^2\right) + h(K, \gamma', T_0) \\
&= 2h(K, \gamma', T_0) + 2a\log(2a) = \tau\ .
\end{aligned}$$

□

Before presenting the regret bound in its most general form, we introduce the following term.

$$\tilde{C}_\epsilon := \max\left(T_\epsilon, e^{\frac{1}{\epsilon}}\right) + \frac{2\alpha\eta^{-1/2} - 2 - \epsilon}{2\alpha\eta^{-1/2} - 1 - \epsilon}\left(\frac{MK^2\left(3 + 2\log\left(d_{\max}^{\gamma} + 1\right)\right)}{\min\{1, 2\alpha\eta^{-1/2} - 1\}\delta\log\eta}\right)^{\frac{1}{2\alpha\eta^{-1/2} - 1 - \epsilon}}\ .$$

Theorem E.5. *For any $\epsilon > 0$, $\gamma' \leq \gamma$ and $\delta > 0$, denote $\Gamma(m, \gamma')$ as the size of the largest clique player m belongs to in $\mathcal{G}_{\gamma'}$, and define,*

$$\begin{aligned}
\hat{C}_\epsilon(\delta) :&= \Delta_{1\max}\sum_{m=1}^{M}\min_{\gamma' \leq \gamma}\left(K^2(\gamma' + 2) + \frac{2\mathcal{D}}{\Gamma(m, \gamma')}\log(\frac{2\mathcal{D}}{\Gamma(m, \gamma')})\right) \\
&\quad + (2C_\epsilon(\delta) + K(3\gamma + 2))M\Delta_{1\max}\ ,
\end{aligned}$$

and,

$$\begin{aligned}
\hat{C}_\epsilon :&= \Delta_{1\max}\sum_{m=1}^{M}\min_{\gamma' \leq \gamma}\left(K^2(\gamma' + 2) + \frac{2\mathcal{D}}{\Gamma(m, \gamma')}\log(\frac{2\mathcal{D}}{\Gamma(m, \gamma')})\right) \\
&\quad + (2\tilde{C}_\epsilon + K(3\gamma + 2))M\Delta_{1\max}\ .
\end{aligned}$$

Let $\chi(\mathcal{G}_\gamma)$ denote the clique covering number of $\mathcal{G}_\gamma$. Then the following holds for Algorithm 2.

(a) For any α such that $\alpha > 0.5\sqrt{\eta}(1+\epsilon)$ and all times T, with probability larger than $1 - \delta$,

$$\mathcal{R}(T) \leq \sum_{j>1}\frac{2\alpha\chi(\mathcal{G}_\gamma)}{\Delta_{1j}}\log T + \hat{C}_\epsilon(\delta)$$

(b) For any α such that $\alpha > 0.5\sqrt{\eta}(2+\epsilon)$ and all times T,

$$\mathbb{E}\mathcal{R}(T) \leq \sum_{j>1} \frac{2\alpha\chi(\mathcal{G}_\gamma)}{\Delta_{1j}} \log T + \hat{C}_\epsilon \ .$$

Proof. First, we observe the following fact. For $t > C_\epsilon(\delta)$ with probability larger than $1-\delta$, no pair $(i,i) \neq (1,1)$ can be drawn by any player. Since every player m recommends an arm i iff $i_m(t) = j_m(t) = i$, messages received at rounds $t > C_\epsilon(\delta) + \gamma$ contain $\mathcal{B}^m_{\text{rec}} \in \{\emptyset, 1\}$.

For any $\gamma'(m) \leq \gamma$ denote by $C^m_{\gamma'(m)}$ as a maximal clique in $\mathcal{G}_{\gamma'(m)}$ that player m belongs to with the maximal size, and denote its size as $\Gamma(m, \gamma'(m))$. The proof for the first part relies on the following observation. Lemma E.3 implies that with probability $1-\delta$, for all such cliques and all players m, times $t > T_0 = C_\epsilon(\delta) + \gamma$ and arm pairs $(i,j) \neq (1,1)$,

$$\sum_{m \in C^m_{\gamma'(m)}} N^{m,T_0}_{ij}(t) \leq D_{ij} \log t + (\gamma'(m)+2)\Gamma(m,\gamma'(m)) \ .$$

Let $\hat{T}^m_\epsilon(\delta)$ be a round such that $\hat{T}^m_\epsilon(\delta) > C_\epsilon(\delta) + \gamma + \frac{K(K-1)}{2}(\gamma'(m)+2) + \frac{\mathcal{D}}{\Gamma(m,\gamma'(m))} \log \hat{T}^m_\epsilon(\delta)$. At round $\hat{T}^m_\epsilon(\delta)$, it holds that,

$$\begin{aligned}
\sum_{m \in C^m_{\gamma'(m)}} N^{m,T_0}_{11}\left(\hat{T}^m_\epsilon(\delta)\right) &= \Gamma(m,\gamma'(m))(\hat{T}^m_\epsilon(\delta) - C_\epsilon(\delta) - \gamma) - \sum_{i<j} \sum_{m \in C^m_{\gamma'(m)}} N^{m,T_0}_{ij}\left(\hat{T}^m_\epsilon(\delta)\right) \\
&\geq \Gamma(m,\gamma'(m))(\hat{T}^m_\epsilon(\delta) - C_\epsilon(\delta) - \gamma) - \sum_{i<j} D_{ij} \log \hat{T}^m_\epsilon(\delta) - \frac{K(K-1)}{2}(\gamma'(m)+2)\Gamma(m,\gamma'(m)) \\
&> 0 \ ,
\end{aligned}$$

where we used the fact that at each round there are $\Gamma(m,\gamma'(m))$ draws in the clique $C^m_{\gamma'(m)}$. Therefore, for at least one player $m' \in C^m_{\gamma'(m)}$ the pair $(1,1)$ is drawn at least once between rounds $C_\epsilon(\delta) + \gamma$ and $\hat{T}^m_\epsilon(\delta)$. We now consider two scenarios.

- If $m = m'$, at the round of this first draw, for all $j > 1$, we have $u^m_{j1} < 0.5$, so that arm 1 is the only arm in the set $\mathcal{C}^m$ and $\mathcal{B}^m = \{1\}$ for this round regardless of what recommendations arrive.

- If $m' \neq m$, a recommendation about arm 1 arrives to player m before round $\hat{T}^m_\epsilon(\delta) + \gamma'(m)$. At that round $\mathcal{B}^m_{\text{rec}} \cap \mathcal{C}^m = \{1\}$ since a recommendation about no other arm is possible, and so it updates $\mathcal{B}^m = 1$.

In both scenarios, player m updates its $\mathcal{B}^m$ set to 1 before round $\hat{T}^m_\epsilon(\delta) + \gamma'(m)$. Since arm 1 remains in $\mathcal{C}^m$ for subsequent rounds and no recommendation other than 1 or $\emptyset$ arrives, the set $\mathcal{B}^m$ never changes going forward - so the first arm drawn after round $\hat{T}^m_\epsilon(\delta) + \gamma'(m)$ by player m is always arm 1.

By Lemma E.4 it holds that for any $\gamma'(m) \leq \gamma$,

$$\hat{T}^m_\epsilon(\delta) \leq 2C_\epsilon(\delta) + 2\gamma + K(K-1)(\gamma'(m)+2) + \frac{2\mathcal{D}}{\Gamma(m,\gamma'(m))} \log(\frac{2\mathcal{D}}{\Gamma(m,\gamma'(m))}) \ ,$$

so with probability $1-\delta$ we decompose the regret with respect to the different cliques in $\mathcal{C}_\gamma$,

$$\begin{aligned}
\mathcal{R}(T) &= \sum_{i\leq j}\sum_{m=1}^{M}\frac{\Delta_{1i}+\Delta_{1j}}{2}N_{ij}^{m}(T) \\
&\leq \sum_{i\leq j}\sum_{m=1}^{M}\frac{\Delta_{1i}+\Delta_{1j}}{2}N_{ij}^{m}(\hat{T}_\epsilon^m(\delta)+\gamma'(m)) \\
&+\sum_{i\leq j}\sum_{C\in\mathcal{C}_\gamma}\sum_{m\in C}\frac{\Delta_{1i}+\Delta_{1j}}{2}N_{ij}^{m,\hat{T}_\epsilon^m(\delta)+\gamma'(m)}(T) \\
&\leq \Delta_{1\max}\sum_{m=1}^{M}(\hat{T}_\epsilon^m(\delta)+\gamma'(m))+\sum_{j>1}\frac{\Delta_{1j}}{2}\sum_{C\in\mathcal{C}_\gamma}\sum_{m\in C}N_{ij}^{m,\hat{T}_\epsilon^m(\delta)+\gamma'(m)}(T) \\
&\leq \Delta_{1\max}\sum_{m=1}^{M}(\hat{T}_\epsilon^m(\delta)+\gamma'(m))+\sum_{j>1}\frac{\Delta_{1j}}{2}\sum_{C\in\mathcal{C}_\gamma}\sum_{m\in C}N_{ij}^{m,C_\epsilon(\delta)}(T) \\
&\leq \Delta_{1\max}\sum_{m=1}^{M}(\hat{T}_\epsilon^m(\delta)+\gamma'(m))+\sum_{j>1}\frac{\Delta_{1j}}{2}\sum_{C\in\mathcal{C}_\gamma}(D_{1j}\log T+(\gamma+2)|C|) \\
&\leq \Delta_{1\max}\sum_{m=1}^{M}\left(K^2(\gamma'(m)+2)+\frac{2\mathcal{D}}{\Gamma(m,\gamma'(m))}\log(\frac{2\mathcal{D}}{\Gamma(m,\gamma'(m))})\right) \\
&+(2C_\epsilon(\delta)+K(3\gamma+2))M\Delta_{1\max}+\sum_{j>1}\frac{\Delta_{1j}}{2}\sum_{C\in\mathcal{C}_\gamma}D_{1j}\log T\ .
\end{aligned}$$

In the derivation above we were free to select any $\gamma'(m)\leq\gamma$ for each player, so for the final bound we get,

$$\begin{aligned}
\mathcal{R}(T) &\leq \Delta_{1\max}\sum_{m=1}^{M}\min_{\gamma'\leq\gamma}\left(K^2(\gamma'+2)+\frac{2\mathcal{D}}{\Gamma(m,\gamma')}\log(\frac{2\mathcal{D}}{\Gamma(m,\gamma')})\right) \\
&+(2C_\epsilon(\delta)+K(3\gamma+2))M\Delta_{1\max}+\sum_{j>1}\frac{\Delta_{1j}}{2}\sum_{C\in\mathcal{C}_\gamma}D_{1j}\log T \\
&\leq \hat{C}_\epsilon(\delta)+\sum_{j>1}\frac{\Delta_{1j}}{2}\chi(G_\gamma)D_{1j}\log T\ .
\end{aligned}$$

To prove the second part, note that the following holds for any random variable X: $\mathbb{E}X=\int_0^1 F_X^{-1}(q)dq$, where $F_X^{-1}(\cdot)$ stands for the quantile function of X. For some invertible function $H(q)$, given that $F_{\mathcal{R}(T)}(H(q))=P\left(\mathcal{R}(T)\leq H(q)\right)>q$, it holds that $F_{\mathcal{R}(T)}(r)>H^{-1}(r)$, so $F_{\mathcal{R}(T)}^{-1}(q)<H(q)$. In our case with $q=1-\delta$, $H(q)$ depends on q only through the first term in the regret containing $\hat{C}_\epsilon(1-q)$. For clarity, we denote,

$$\hat{C}_\epsilon(\delta)=\max\left(A,\left(\frac{B}{\delta}\right)^{\frac{1}{2\alpha\eta^{-1/2}-1-\epsilon}}\right)\ ,$$

where A,B are some terms that are not functions of δ. For $2\alpha\eta^{-1/2}>2+\epsilon$, it holds that,

$$\int_0^1\left(\frac{1}{1-q}\right)^{\frac{1}{2\alpha\eta^{-1/2}-1-\epsilon}}dq=\int_0^1\left(\frac{1}{\delta}\right)^{\frac{1}{2\alpha\eta^{-1/2}-1-\epsilon}}d\delta=\frac{2\alpha\eta^{-1/2}-2-\epsilon}{2\alpha\eta^{-1/2}-1-\epsilon}\ .$$

Putting it all together we have that,

$$\mathbb{E}\mathcal{R}(T)\leq\hat{C}_\epsilon+\sum_{j>1}\frac{\Delta_{1j}}{2}\chi(\mathcal{G}_\gamma)D_{1j}\log T\ .$$

□

The derivation of the proof for Theorem 5.2 in the paper follows a similar process, assuming $\alpha > 1.4, \epsilon = e^{-1}, \eta = 1.3$ and utilizing $C(\delta)$. For a graph with complete communication $\gamma = D$, the non-leading term in the regret becomes,

$$\hat{C} := \Delta_{1\max} \sum_{m=1}^{M} \min_{\gamma' \leq D} \left(K^2(\gamma'+2) + \frac{2\mathcal{D}}{\Gamma(m,\gamma')} \log 2\mathcal{D} \right) + (2\tilde{C} + K(3D+2))M\Delta_{1\max} \; .$$

Since a minimization is involved in the calculation of this term, we can observe the two extreme cases where $\gamma' = 0$,

$$\begin{aligned} \hat{C} &\leq \Delta_{1\max} \sum_{m=1}^{M} \left(2K^2 + 2\mathcal{D} \log 2\mathcal{D} \right) + (2\tilde{C} + K(3D+2))M\Delta_{1\max} \\ &= O(MKD + MK^2 \log K / \Delta_{1\min}^2) \; , \end{aligned}$$

and $\gamma' = D$,

$$\begin{aligned} \hat{C} &\leq \Delta_{1\max} \sum_{m=1}^{M} \left(K^2(D+2) + \frac{2\mathcal{D}}{M} \log 2\mathcal{D} \right) + (2\tilde{C} + K(3D+2))M\Delta_{1\max} \\ &= O(MK^2D + K^2 \log K / \Delta_{1\min}^2) \; . \end{aligned}$$

Since both must hold, we have that,

$$\hat{C} = O\left(MKD + \frac{K^2 \log K}{\Delta_{1\min}^2} + MK^2 \min\left\{ \frac{\log K}{\Delta_{1\min}^2}, D \right\} \right) \; .$$

F MORE EXPERIMENTS

In this section, we provide some more details about the experimental results. The code used to create the figures in this paper, as well as the Sushi and Irish datasets and all preference matrices, is available at `https://github.com/ravehor92/Multi-Player-Approaches-for-Dueling-Bandits`. The Irish and Sushi datasets were obtained using the PrefLib dataset library (Mattei and Walsh, 2013). Please make sure to read the 'README.md' file before accessing the code. For the figures in the main paper, we use $\alpha = 3$ for RUCB as both the base algorithm for Algorithm 1 and in Algorithm 2. Regarding RMED2FH, we use $\alpha = 3$ and $f(K) = 0.3K^{1.01}$ following Komiyama et al. (2015). For FYLBB algorithms, we utilize the simple leader election algorithm from Appendix D.1. The VDB algorithm is run with a time-dependent learning rate $\eta_t = 4/\sqrt{t}$ as indicated in Saha and Gaillard (2022), and the emergent optimization problem at each round is solved using the cvxpy library in Python. All regrets are averaged over 200 independent runs, and the error bars indicate the standard deviation over the runs as calculated using numpy.std. Figures in the main paper use $\gamma = D$. Additionally, the horizontal axis in all plots is scaled logarithmically with T to highlight the $\log T$ dependence of the regret in the experiments. The vertical axis is either linear or logarithmic, chosen to improve plot visibility. For consistency, subfigures within the same figure share identical axis properties.

Our experiments were conducted using Python 3.6 on a system featuring a 2.2 GHz Intel Xeon E5-2698 v4 CPU with 15 cores and 512GB of DDR4 RAM operating at 2133 MHz. All experiments were completed within several hours.

The large variance in Figure 4c originates from the RMED2FH algorithm used by the leader, as asymptotically optimal algorithms explore less and tend to have larger variance. In addition, since the exploration in Figure 4c originates solely from the leader and the followers exploit its candidate arm, this effect is amplified when there are more players. In contrast, exploration in Figure 4a is done cooperatively so the variance is reduced.

Figure 5 illustrates the group regret per round for the Sushi dataset, employing the message-passing RUCB algorithm with 100 players on a cyclic communication graph. The parameter γ is varied to observe its impact on the regret. As γ decreases, the slope in the asymptotic region increases, and notably, the case where $\gamma = D$ aligns its slope closest to the lower bound. This alignment corresponds consistently with the findings outlined in Theorem 5.2.

In Figure 6, we present the regret per round for the Sushi dataset, with M=100 players arranged on a star graph. We compare the performance of the Message Passing RUCB algorithm, as described in Section 5, with and without CW recommendations. Setting the delay to $\gamma = 1$ allows us to illustrate the advantage of CW

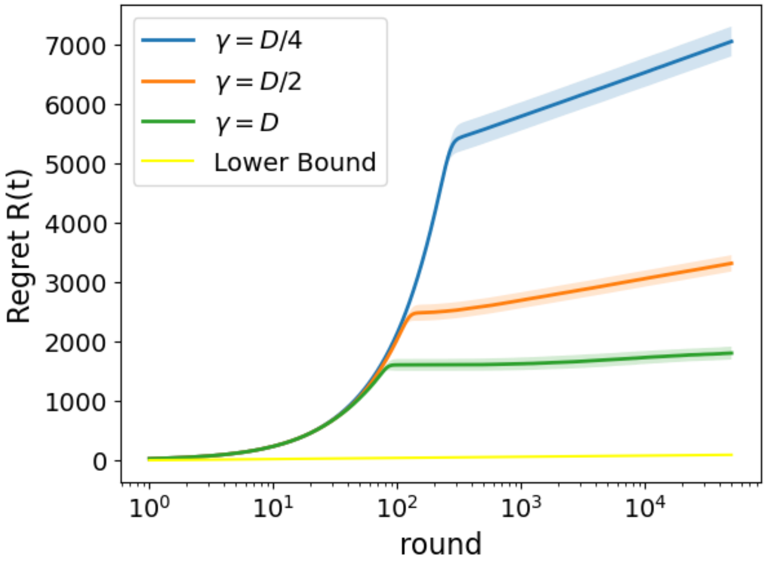


Figure 5: Regret for the Sushi dataset, with the message passing RUCB algorithm and M=100 players. We use a cyclic communication graph and various values for the decay parameter γ.

recommendations in quickly identifying the CW. In this scenario, the central player has immediate access to all other players' observations, while the peripheral players can only utilize their own observations and those of the central player. Consequently, we anticipate that the central player will identify the CW faster than the peripheral players. Once this occurs a CW recommendation will benefit all peripheral players by helping them identify the CW and focus on exploring pairs of arms where at least one arm is the CW. This approach helps address the inherent challenge in the multi-player dueling bandit setting, as discussed in Section 1, which involves exploring non-CW pairs of arms.

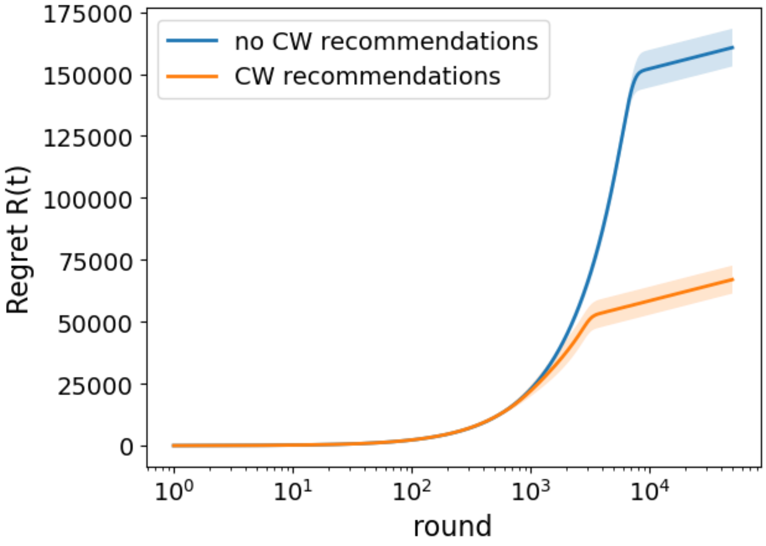


Figure 6: Regret for the Sushi dataset, with M=100 players on a star graph and $\gamma = 1$. We use the message passing RUCB algorithm from Section 5, as well as a version without CW recommendations.

Both versions of the Message Passing UCB exhibit similar asymptotic performance. However, as shown in

Figure 6, utilizing CW recommendations leads to a quicker identification of the CW and subsequently results in a much smaller regret. This effect is particularly pronounced in experimental settings where communication graphs have $\gamma < D$.