Collision Resolution in Multi-player Bandits Without Observing Collision Information

Eleni Nisioti [1] **Nikolaos Thomos** [2] **Borris Bellalta** [3] **Anders Jonsson** [3]

Abstract

The absence of collision information in Multi-player Multi-armed bandits (MMABs) renders arm availabilities partially observable, impeding the design of algorithms with regret guarantees that do not allow inter-player communication. In this work, we propose a collision resolution (CR) mechanism for MMABs inspired from sequential interference mechanisms employed in communication protocols. In the general case, our collision resolution mechanism assumes that players can pull multiple arms during the exploration phase. We, thus, propose a novel MMAB model that captures this while still considering strictly bandit feedback and single-pulls during the exploitation phase. We theoretically analyze the CR mechanism using tools from information theory in order to prove the existence of an upper bound on the probability of its failure that decreases at a rate exponential in the number of players.

1. Introduction

In a multi-armed bandit (MAB), a player sequentially interacts with a finite set $\mathcal{K} = \{1, ..., K\}$ of arms that incur rewards following unknown probability distributions with the aim of maximizing the reward it accrues by the end of the problem horizon T. Multi-player multi-armed bandits (MMABs) generalize this framework to the case where $M \geq 2$ players compete for this set of arms. Interest in such algorithms has recently been reignited (Boursier and Perchet, 2018; Rosenski et al., 2016; Besson and Kaufmann, 2017; Boursier et al., 2020; Shi and Shen, 2020) primarily due to their applicability in dynamic spectrum access (Zhao and Sadler, 2007), where channels are viewed as arms and communication devices as players searching for the optimal assignment in an online and decentralized manner.

In this work, we consider problem settings where the rewards of an arm k follows a Bernoulli distribution with mean μ_k, there is no communication among players, and their number, M, is unknown a priori. When more than one player simultaneously pulls the same arm, a collision occurs and players involved observe zero reward. In our setting, often referred to as the *no-sensing* setting, we assume that only rewards are observable, rendering the bandit feedback partially observable.

Our work is motivated by the following question: *is it possible to theoretically guarantee that a player can compute unbiased mean estimations of arms availabilities' without communication and without observing collisions?* We answer this question in the affirmative by introducing a collision resolution (CR) mechanism. Due to space limitations, we focus on the description of this mechanism and its theoretical analysis. To illustrate how our mechanism can be employed in practice, we include the description of a bandit algorithm that makes use of it, Dynamic CR-UCB (DYN-CR-UCB) in Appendix 5 and present numerical experiments. DYN-CR-UCB improves upon previous works that address the same question, which provided unrealistic regret bounds (Lugosi and Mehrabian, 2018) or exploited collisions to communicate information indirectly instead of implicitly resolving them (Boursier and Perchet, 2018; Shi and Shen, 2020).

The CR mechanism is inspired by the realization that the no-sensing reward model gives rise to a learning process that can be studied using the information-theoretic AND-OR tree analysis, originally proposed by Luby et al. (1998). Our proposed CR mechanism dictates how players should behave in a CR round. The theoretical analysis of the probability of failure to resolve collisions in a given round allows us to upper bound its duration. Our theoretical analysis of the CR mechanism is based on the evolution of random processes on bipartite graphs, which represent a resource allocation problem using a set of nodes for players and another set of nodes for resources (Luby et al., 1998). The analysis unfolds in two steps. First, the probability of failure of a CR round is computed for asymptotic settings ($M, K \to \infty$) and the minimum duration of a round for a target probability is com-

[1]Flowers Team, Inria and Ensta ParisTech [2]University of Essex [3]Pompeu Fabra University. Correspondence to: Eleni Nisioti <eleni.nisioti@inria.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

puted. This analysis is simple because, in an asymptotic setting, the probability of failure evolves independently for each player. Then, we compute a concentration inequality that bounds the deviation of the performance of graphs with a finite number of nodes from the asymptotic one.

Our CR mechanism requires the introduction of a bandit model that has not been previously considered in the literature. During the exploitation phase, our model is identical to the classical no-sensing MMAB model used by Lugosi and Mehrabian (2018); Besson and Kaufmann (2017); Rosenski et al. (2016). During the exploration phase, however, a player has the ability to simultaneously pull multiple arms and observe their rewards. Until now, bandit models have been either single-pull or multiple-pull (Agrawal et al., 1990). Our bandit model is a hybrid appropriate for modeling collision resolution mechanisms classically employed in resource allocation tasks with random access, as is our considered MMAB setting. The application that we consider, i.e., cognitive radio, only requires that agents exploit a single arm; pulling multiple arms is a mechanism for resolving collisions during exploration. We should also note that the setting considers strictly bandit feedback, and the multiple pulls cannot be considered as side information considered in other works (Degenne et al., 2018), as they are equally amenable to collisions.

2. Related Work

The no-sensing reward model has not been extensively studied in the MMAB literature, arguably due to the difficulty of theoretically analyzing it and designing algorithms with practical sample complexity. In the family of *Selfish* bandit algorithms (Besson and Kaufmann, 2017; Bonnefoi et al., 2018), partial observability is ignored, leading to a loss of regret guarantees as the collected samples are biased due to collisions. One of the few attempts to address this problem is the algorithm introduced by Lugosi and Mehrabian (2018), where players independently compute unbiased estimates of the means of arms availabilities' by scaling empirical means with the probability of collision. In our work, the CR mechanism offers an alternative way of calculating unbiased estimates with significantly reduced sample complexity.

An approach orthogonal to ours is the exploitation of collisions to communicate statistical information indirectly (Boursier and Perchet, 2018; Shi and Shen, 2020). Under the assumption that all the players start learning simultaneously, referred to as a *synchronized* setting, it is possible to achieve bounds similar to those of a centralized setting (Boursier and Perchet, 2018). Recently, EC-SIC (Shi et al., 2020) improved upon SIC-MMAB2 (Boursier and Perchet, 2018) in the no-sensing setting by introducing channel coding. Although the observation that indirect communication can help bridge the gap between centralized and decentralize settings is inspiring, it comes at the cost of requiring synchronization and communication time that increases sample complexity.

An important trait of a MMAB algorithm is whether it is dynamic, i.e., whether regret guarantees can be derived when players arrive at different time steps. SIC-MMAB (Boursier and Perchet, 2018) and EC-SIC (Shi and Shen, 2020) require that all the players start learning together in order to acquire the correct statistics. In contrast, inherently dynamic algorithms, such as Selfish (Besson and Kaufmann, 2017), DYN-MMAB (Boursier and Perchet, 2018) and our proposed algorithm, DYN-CR-UCB, naturally deal with dynamic settings.

The analysis of iterative message passing algorithms on graphical models has a long history (Liva, 2011; Luby et al., 2001; Luby et al., 1997) and has served as the basis for the analysis of belief propagation algorithms, employed in a variety of applications, such as collision resolution in Medium Access Control protocols (Liva, 2011) and belief propagation decoders in channel codes (Luby et al., 2001). Important steps in this analysis have been the introduction of density evolution for describing the evolution of messages in asymptotic settings and the derivation of concentration bounds characterizing the performance for finite lengths (Richardson and Urbanke, 2001). To the best of our knowledge, our work is the first attempt to transfer this analysis to bandits, which differ from previously studied resource allocation problems in that resources are not always available and there exists no centralized point of control.

3. Bandit model

We consider a K-armed bandit, where each arm is characterized by its availability which follows a Bernoulli distribution with mean μ_k. We denote by $y_k(t)$ the independent and identically distributed (iid) random variable associated with each arm, satisfying $P\big((y_k(t)) = 1\big) = \mu_k$, and refer to it as the availability of the arm. At each time step t, each one of the M players chooses a subset $\mathcal{D}$ of the K arms to pull, an action we denote as $\vec{a}_m(t)$. Upon being pulled, an arm returns a reward of one if it is available, i.e., $y_k(t) = 1$, and only one player pulled it. Differently, the returned reward is zero. The reward model is formally:

$$r_m(t) = \vee\{y_k(t)(1 - n_k(t)), \ \ \forall k \in \vec{a}_m(t)\}, \qquad (1)$$

where $\vee$ denotes that the logical-or operation is applied on the vector of rewards collected by all arms pulled by player m and $n_{a_m}(t)$ indicates whether players collided on the arm a_m. A player observes a set of rewards for each arm pull in its subset and their final reward is the OR function of the observed rewards.

The sampling process is repeated for T steps, where T is termed the problem horizon, fixed and known in advance.

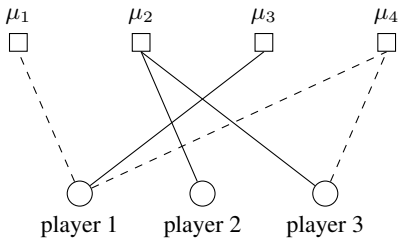

Figure 1. A bipartite-graph representation of a MMAB with 4 arms and 3 players. Until now, algorithms for MMABs have considered that players pull a single arm (solid lines). In this work, players can pull multiple arms simultaneously (solid and dashed lines).

The players' objective is to minimize the expected cumulative regret at the end of the horizon, defined as:

$$R(T) = T\sum_{k=1}^{M}\mu_k^* - \sum_{t=1}^{T}\sum_{m=1}^{M} r_m(t), \tag{2}$$

where μ_k^* is the mean availability of an M-best arm, i.e., an arm belonging to the set of arms with the M highest means, which we denote by $\mathcal{M}^*$.

In Figure 1, we illustrate a single time step in a bandit setting with $K = 4$ arms and $M = 3$ players. We adopt the common practice of representing the problem as a bipartite graph with two sets of nodes: lower nodes represent players and upper nodes represent arms.

4. The collision resolution mechanism

The main novelty in our work is the realization that the reward model presented in (1) can be studied using the AND-OR tree analysis, which describes the evolution of random processes on graphs (Luby et al., 1998). Let us revisit the example of Figure 1 to describe the CR mechanism. We assume that arms are always available, i.e., $\mu_k = 1, \forall k \in \mathcal{K}$, which allows us to decouple the CR mechanism from the sampling process. If we only consider the solid lines, then each player pulls a single arm, with the end result of players 2 and 3 colliding and observing a reward of zero, regardless of the availability. If we now consider both dashed and solid lines, players 1, 2 and 3 pull three, one and two arms, respectively. At first sight, this strategy looks counter-intuitive; players experience more collisions than previously, while the end result remains the same: only player 1 observes a reward of one. This, however, changes if we employ the following CR mechanism: each player repeats its actions (i.e., pulling the same subset of arms) until it observes a reward of one from one of them. When this happens, the player stops pulling the other arms and keeps pulling the arm that gave the reward of one. If more than one arm returns a positive reward, one of them is selected randomly. This process lasts for $I_{\max}$ iterations, which form a single CR round. In our example, this mechanism will lead to a collision-free assignment within 3 time steps (remember that $\mu_k = 1, \forall k \in \mathcal{K}$). First, player 1 receives a reward of one from arm 1, and, hence, stops pulling arms 3 and 4. Then, player 3 receives a reward of one from arm 4 and stops pulling arm 2, and finally, player 2 receives a reward of one from arm 2. For the remaining iterations, up to $I_{\max}$, the three players continue pulling the arms that gave a reward of one and collect unbiased observations. We should note that we kept the size of the problem studied in this example small to provide an intuitive explanation of the CR mechanism, which was however conceived for asymptotic settings ($M, K \to \infty$), where its probability of failure is arbitrarily small (Luby et al., 1998).

To make the discussion more formal we introduce some additional notation. Let $\Lambda_m(t)$ be the random variable denoting the number of arms player m pulls at time step t, which follows a multinomial probability distribution with coefficients $[\Lambda_1, \ldots, \Lambda_D]$, where D is the maximum number of simultaneous pulls allowed to a player and Λ_d denotes the probability of pulling d arms. We refer to $\Lambda_m(x)$ as a degree distribution and, as is common in the analysis of bipartite graphs, describe it using the generating function $\Lambda_m(x) = \sum_{d=1}^{D} L_d x^d$. We consider anonymous settings, i.e., $\Lambda_m(x) = \Lambda(x),\ \forall m \in \mathcal{M}$.

Algorithm 1 contains the pseudocode of a CR round. It requires as inputs the set of arms $\mathcal{K}$, the degree distribution $\Lambda(x)$ and the duration of a CR round, $I_{\max}$. All algorithms in our work are presented from the perspective of a single player, as they are run in parallel by all players. At the beginning of a round, the player selects $\mathcal{D}$, a subset of arms to pull (Lines 1-2) whose size is determined by sampling $\Lambda(x)$. When a collision-free arm is found (Line 7), all other arms are removed from the set $\mathcal{D}$, which thus becomes a singleton, and its estimate is updated (Lines 12-13).

An important step in the theoretical analysis of our proposed algorithm is to determine the conditions under which the CR mechanism succeeds with high probability. A round completes successfully for a player when it has found a collision-free arm by the end of it. The following theorem states that the probability of failure of the CR mechanism is upper bounded by a value that depends on the number of players and the probabilistic structure of the bipartite graph.

Theorem 4.1. *Assume that a CR round of I_{max} iterations takes place among M players accessing K arms and the degree distribution is $\Lambda(x)$. Then, a CR round fails for a player with probability at most:*

$$p_f = q_{I_{max}} + \sqrt{\frac{-\ln(\delta_4)}{\eta M}},$$

where $q_{I_{max}}$ is the probability of failure in an asymptotic setting ($M, K \to \infty$) at the end of the CR round, δ_4 is

Algorithm 1 CR round

Decide d, the number of arms to pull simultaneously by sampling $\Lambda(x)$
Form a random subset $\mathcal{D}$, by randomly choosing d out of the K arms
free = *False*
for $\tau \in \{1, \dots, I_{\max}\}$ **do**
  Pull all arms in $\mathcal{D}$,
  Observe rewards $r_{\mathcal{D}}$,
  if $\exists i : r_i == 1$ and not free **then**
    Remove all elements except i from $\mathcal{D}$ {Detect collision-free arm}
    free = *True*
  end if
  if free **then**
    $S_{\mathcal{D}} = S_{\mathcal{D}} + r_{\mathcal{D}}$ {Update sum of rewards for collision-free arm}
    $T_{\mathcal{D}} = T_{\mathcal{D}} + 1$ {Update number of pulls for collision-free arm}
  end if
end for
Return free, $\mathcal{D}$

the probability that the second term on the right-hand side has been under-estimated and η is a constant whose value depends on the structure of the graph and is given in Lemma B.3 in Appendix B.

Proof (Sketch). Our proof requires results found in different works from the field of information theory (Luby et al., 2001; Sipser and Spielman, 1996; Richardson et al., 2001; Liva, 2011; Luby et al., 1998; 1997). We, therefore, deemed it necessary for the completeness of our work to gather these results and adjust them to our problem setting. In Appendix A, we provide a general description of bipartite graphs and present Lemma A.1, which bounds the probability that a bipartite graph of finite size does not have a tree structure, and in Appendix B, we present the analysis of the CR mechanism. The first step of the proof, in Lemma B.1, is to derive the condition under which the probability of failure of the CR mechanism is monotonically decreasing with each iteration in a round. This condition, referred to as the stability condition, is derived assuming that M is infinite, which simplifies the analysis as it guarantees that the graph is cycle-free, meaning that the probability of failure evolves independently for each player. Thus, we can compute the duration of a CR round, $I_{\max}$, based on a target probability of error for the CR mechanism, $q_{I_{\max}}$. Note that we have slightly modified the existing analysis to take into account the effect of arms availabilities', i.e., μ_k, have on the calculation of $I_{\max}$, as the original proof considered a setting with $\mu_k = 1, \forall k \in \mathcal{K}$. In order to transfer the analysis to settings with finite M and K, where cycles may appear on the bipartite graph, in Lemma B.3 we formulate the process of resolving collisions as a martingale and derive a concentration inequality that describes how the probability of failure diverges from its asymptotic expectation. We make use of Lemma B.3 by setting the right-hand side of (11) to be equal to δ_4, which leads to the value of $\alpha = \sqrt{\frac{-\ln(\delta_4)}{\eta M}}$. □

We should note that the bound appearing in Theorem 4.1 is not valid unconditionally. In particular, Lemma B.3 introduces a condition on the minimum number of players M, i.e., $M > 2\gamma/\alpha$, where γ is a constant that depends on the probabilistic structure of the bipartite graph and is defined in Lemma A.1, and α was defined above, for the result to be valid. In order to derive this condition, the analysis of Richardson and Urbanke (2001) makes a very conservative estimation which relies on the assumption that cycles of any length in the bipartite graph can lead to failure of the CR mechanism, by requiring that $l = I_{\max}$ in the estimation of γ. While it has been empirically observed that only cycles of very small length affect the performance of random processes on graphs (Richardson and Urbanke, 2001), this conjecture remains to be theoretically proven.

5. The DYN-CR-UCB algorithm

In the following, we describe an algorithm that employs our CR mechanism in a dynamic setting considered in previous works Boursier and Perchet (2018) where players arriving at different time steps $\tau_m \in \{0, \dots, T-1\}$, where τ_m is unknown to all. We denote the learning horizon of player m by T_m. A player knows the time elapsed since joining the network and observes a common clock with period $I_{\max}$ and can be in one of the two phases: (i) in the exploration phase, the player is employing the CR mechanism, as it was described in Section 4, and experiences CR rounds of equal duration $I_{\max}$; (ii) in the exploitation phase, the player is pulling a single arm until the end of the horizon..

During the exploration phase, a player computes unbiased estimates $\hat{\mu}_k$ for all arms availabilities' and a confidence bound $B_t = 2\sqrt{\frac{\log(T_m)}{t}}$. Thus, it knows that the true mean of the availability of an arm lies with high certainty in the range $[\hat{\mu}_k - B_k, \hat{\mu}_k + B_k]$. The player keeps an initially empty *preferences* list, $\vec{p}$ and inserts an arm in it once it detects that its lower bound is higher than the upper bound of all other arms in the list. The player will exploit an arm in $\vec{p}$ as soon as it gives a positive reward. As each player employs the UCB algorithm with confidence bound $B_k(t) = \sqrt{\frac{\log(T_m)}{T_k}}$, using Hoeffding's inequality, we can prove that:

$$\mathbb{P}[|\hat{\mu}_k - \mu_k| > B_k] \leq 4/T_m^2, \tag{3}$$

which suggests that all players will acquire a correct esti-

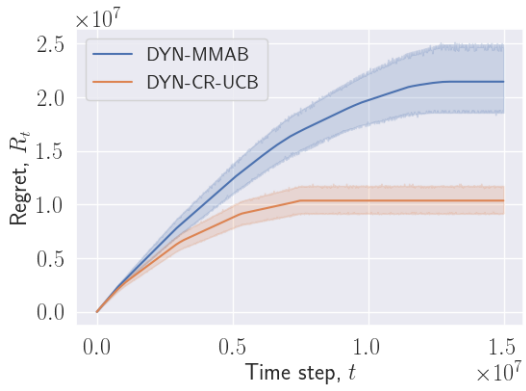

Figure 2. Cumulative regret achieved by DYN-CR-UCB and DYN-MMAB in a dynamic network with $K = 10$, $M = 5$ users arriving at different time steps and $\Delta = 0.06$.

mate of all free arms at the end of their individual horizon T_m with probability $1-\mathcal{O}(1/(T_m)^2)$. In addition, we know that a sub-optimal arm is detected within $K \log T_m/\Delta_k^2$ time steps, where $\Delta_k = \min_{i=1,\cdots,k} |\mu_i - \mu_{i+1}|$ indicates the difficulty of ranking an arm.

An important element of DYN-CR-UCB is how the detection of exploited arms happens. In contrast to DYN-MMAB, where players sample arms randomly and cannot discern between an occupied and an unavailable arm, players in our setting are employing the CR mechanism. Due to this, they know that, at the end of a CR round, at least one arm will be observed as available, provided that the CR round completes successfully. Thus, the probability of an arm being unavailable even if it is not exploited is equal to p_f and is, thus, independent of its mean availability. This significantly reduces the sample complexity compared to DYN-MMAB.

The number of consecutive rounds of observing no rewards from an arm required to declare it as occupied, L, needs to be high enough to guarantee that it is not falsely detected as occupied and low enough to ensure that detection does not incur unnecessary regret. By setting $L \geq 2\log T_m/(1-p_f)$, we ensure that the probability of observing L successive rounds with all-zero rewards is smaller than $\frac{1}{(T_m)^2}$, due to the inequality $(1-p_f)^L \leq e^{-L(1-p_f)}$. In order to prove that a player will pull an arm L times with probability $1/T_m^2$, we make use of the Hoeffding bound of the binomial distribution which takes the value one with probability equal to the probability of sampling arm k, denoted as $p_k = \sum_{l=0}^{d} \Lambda_d l/K$. This leads to $L_2 = \frac{Lp_k \pm L^2((p_k-1)p_f+\log T_m)}{p_k-\log T_m}$. Thus, if a player occupies an arm at time step $t_0+\tau_j$, then it is correctly detected as occupied within $\mathcal{O}(I_{\max}L_2)+\tau_j$ steps, where we have taken into account that a round lasts for $I_{\max}$ iterations.

By making use of Lemma 10 presented by Boursier and Perchet (2018) and the preceding discussion we derive the following regret bound for DYN-CR-UCB:

Theorem 5.1. *In the dynamic setting, the regret of DYN-CR-UCB is upper bounded as follows:*

$$\mathbb{E}[R_T] \leq \frac{MK\log T}{\Delta^2(M)} + MI_{max}L_2,$$

where $\bar{\Delta}^2(M) = \min_{i=1,\ldots,M} |\mu_i - \mu_{i+1}|$, $p_k = \sum_{l=0}^{d} \Lambda_d l/K$ *and*

$$L_2 = \min\Big\{\frac{(Lp_k + L^2((p_k-1)p_f+\log T_m)}{p_k-\log T_m}, \frac{Lp_k - L^2((p_k-1)p_f+\log T_m)}{p_k-\log T_m}\Big)\Big\} \tag{4}$$

5.1. Simulations

We consider a problem setting with $K = 10$ arms, a horizon $T = 15\cdot 10^6$, minimum distance $\Delta = 0.06$ and 5 players arriving at time steps randomly sampled in $[0.05T, 0.575T]$, with the first player always arriving at the first time step. In Figure 2, we compare the performance of DYN-CR-UCB, with that of DYN-MMAB (Boursier and Perchet, 2018). We observe that players using our proposed algorithm DYN-CR-UCB, find an optimal arm significantly quicker than players employing DYN-MMAB and exhibit lower variance.

6. Discussion

We have presented a collision resolution mechanism for multi-player bandits in the no-sensing setting. Our main motivation has been to show that the problem of exploration under partial observability in MMABs can be efficiently addressed by appropriately orchestrating the learning process. Our work is an important step towards designing algorithms with improved regret bounds in the no-sensing setting based on the intuition that collisions can be resolved despite the absence of sensing information. Crucially, the performance of the CR mechanism improves with the number of players, while the communication of statistics (Boursier and Perchet, 2018; Shi et al., 2020) introduces a large overload.

From an application perspective, our solution has some limitations. First, similarly to many recent works in MMABs, fairness is not taken into account. Our approach satisfies a weaker notion of fairness; players have equal chances of finding the best arm across independent trials. Furthermore, the employed mechanism of simultaneously pulling multiple arms is associated with slightly increased complexity. However, such an increase is affordable in multiple access schemes used in wireless communications. To give a clearer picture of the introduced complexity, the number of multiple pulls ranges from 1 to 3 for small networks (less than 200 users) and can reach up to 8 for larger networks.

References

Agrawal, R., Hegde, M., and Teneketzis, D. (1990). Multi-armed bandit problems with multiple plays and switching cost. *Stochastics An International Journal of Probability and Stochastic Processes*, 29:437–459.

Besson, L. and Kaufmann, E. (2017). Multi-player bandits revisited.

Bonnefoi, R., Besson, L., Moy, C., Kaufmann, E., and Palicot, J. (2018). Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings.

Boursier, E., Kaufmann, E., Mehrabian, A., and Perchet, V. (2020). A practical algorithm for multiplayer bandits when arm means vary among players.

Boursier, E. and Perchet, V. (2018). SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. *CoRR*, abs/1809.08151.

Degenne, R., Garcelon, E., and Perchet, V. (2018). Bandits with side observations: Bounded vs. logarithmic regret.

Liva, G. (2011). Graph-based analysis and optimization of contention resolution diversity slotted aloha. *IEEE Transactions on Communications*, 59(2):477–487.

Luby, M. G., Mitzenmacher, M., and Shokrollahi, M. A. (1998). Analysis of random processes via and-or tree evaluation. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '98, page 364–373, USA. Society for Industrial and Applied Mathematics.

Luby, M. G., Mitzenmacher, M., Shokrollahi, M. A., and Spielman, D. A. (2001). Improved low-density parity-check codes using irregular graphs. *IEEE Transactions on Information Theory*, 47(2):585–598.

Luby, M. G., Mitzenmacher, M., Shokrollahi, M. A., Spielman, D. A., and Stemann, V. (1997). Practical loss-resilient codes. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '97, page 150–159, New York, NY, USA. Association for Computing Machinery.

Lugosi, G. and Mehrabian, A. (2018). Multiplayer bandits without observing collision information. *CoRR*, abs/1808.08416.

Richardson, T. J., Shokrollahi, M. A., and Urbanke, R. L. (2001). Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory*, 47(2):619–637.

Richardson, T. J. and Urbanke, R. L. (2001). The capacity of low-density parity-check codes under message-passing decoding. *IEEE Transactions on Information Theory*, 47(2):599–618.

Rosenski, J., Shamir, O., and Szlak, L. (2016). Multi-player bandits: A musical chairs approach. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 155–163. JMLR.org.

Shi, C. and Shen, C. (2020). On no-sensing adversarial multi-player multi-armed bandits with collision communications.

Shi, C., Xiong, W., Shen, C., and Yang, J. (2020). Decentralized multi-player multi-armed bandits with no collision information. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1519–1528. PMLR.

Sipser, M. and Spielman, D. A. (1996). Expander codes. *IEEE Transactions on Information Theory*, 42(6):1710–1722.

Zhao, Q. and Sadler, B. M. (2007). A survey of dynamic spectrum access. *IEEE Signal Processing Magazine*, 24(3):79–89.

A. Useful Properties of Bipartite Graphs

This appendix includes results related to bipartite graphs that are useful for analyzing the CR mechanism.

We denote a bipartite graph describing a problem setting of M players and K channels as $G(\mathcal{M}, \mathcal{K}, \mathcal{E})$, where $\mathcal{E}$ is the set of edges representing arms pulled by players at a given time step. In our analysis, we refer to nodes representing players as player nodes (PNs), nodes representing arms as arm nodes (ANs) and denote an edge between player m and arm k as $\vec{e} = (m, k)$. An example of a bipartite graph is presented in Figure 3.

Players pull random subsets of arms, with the size of the subset being determined by sampling the degree distribution $\Lambda(x) = \sum_{l=1}^{D} \Lambda_l x^l$, where D is the maximum number of pulls allowed to a player. We denote an ensemble of bipartite graphs as $\mathcal{G}(M, K, \Lambda(x))$, i.e., the family of bipartite graphs that can be generated using this random process. From the perspective of ANs, $\Psi(x) = \sum_{l=1}^{D} \Psi_l x^l$ is the distribution describing the number of pulls on each arm. Thus, the average number of pulls for a player is $\bar{\Lambda} = \sum_{l=1}^{D} l\Lambda_l$ and, equivalently for an arm, $\bar{P} = \sum_{l=1}^{D} lP_l$. This leads to the following relationship for the load of the network: $L = M/K = \Psi'(1)/\Lambda'(1)$. In addition to the $\Lambda(x)$ and $P(x)$ degree distributions, which we term as node-perspective, we also refer to the edge-perspective degree distributions $\lambda(x) = \sum_{l=2}^{D} \lambda_l x^{l-1}$ ($\rho(x) = \sum_{l=2}^{D} \rho_l x^{l-1}$), where $\lambda_l (\rho_l)$ denotes the percentage of edges that are connected to a PN (AN) of degree l.

An important trait of our theoretical analysis is that it concerns randomly built graphs. As a result, the actual connections between PNs and ANs cannot be known in advance and vary for different CR rounds. In order to analyze the performance of the CR mechanism, and thus the regret of DYN-CR-UCB, we need to ensure that the performance of a given graph is close to that of its ensemble. This is termed the *concentration* property of the ensemble, and will be proven in Appendix B for our setting.

An important concept in our analysis is that of a sub-graph, $G_{\vec{e}}^{2l}$, which is obtained by the following process: choose an edge $\vec{e} = (m, k)$ uniformly at random from among all edges of a bipartite graph $G(\mathcal{M}, \mathcal{K}, \mathcal{E})$, and then consider the sub-graph induced by the upper node m and all its neighbors within distance $2l$ after deleting the edge (m, k). Sub-graphs are useful because they help us describe how each step of the CR mechanism affects the structure of the original bipartite graph. An alternative way to describe a sub-graph is through the neighborhood $\vec{\mathcal{N}}_e^{2l}$, which is the set of all nodes and edges included in the corresponding sub-graph $G_{\vec{e}}^{2l}$. Figure 4 presents a $G_{\vec{e}}^{2}$ sub-graph induced for the edge $\vec{e} = (5, 4)$. As the nodes in the sub-graph are distinct (there are no loops), the sub-graph is tree-like.

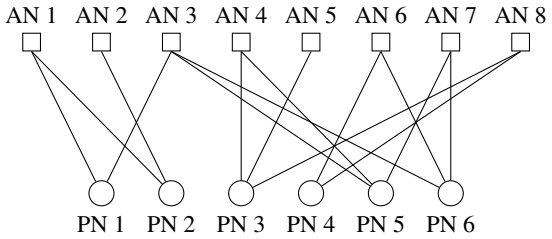

Figure 3. Illustration of a bipartite graph where player nodes' (PNs) degrees are either 2 or 3.

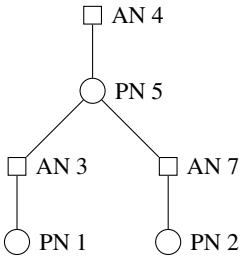

Figure 4. The induced sub-graph for edge (5, 4) and $l = 1$. This sub-graph is tree-like, because no node appears twice.

As we will see in the analysis of the CR mechanism, a tree-like structure is essential for ensuring that all players manage to find a collision-free arm. Lemma A.1 proves that the probability that a sub-graph is not tree-like is negligible for a large enough number of players M. We have derived it by extending existing analysis (Richardson and Urbanke, 2001, Appendix A), which concerned regular bipartite graphs.

Lemma A.1. *Consider a randomly constructed graph G. Let $G_{\vec{e}}^{2l^*}$ be the sub-graph of fixed length $2l^*$ for a given edge $\vec{e}$. Then, for some constant γ:*

$$\mathbb{P}(\mathcal{N}_e^{2l^*} \textit{ is not tree-like}) \leq \frac{\gamma}{M}$$

Proof. Denote with $\Lambda_{\max}$ the maximum degree of a PN and $P_{\max}$ the average degree of an AN. (Note that in other parts of the paper we also refer to $\Lambda_{\max}$ as D.) Under the assumption that the sub-graph is tree-like, the number of PNs in the sub-graph is:

$$M_{l^*} = \sum_{i=0}^{l^*} (\Lambda_{\max} - 1)^i (P_{\max} - 1)^i \tag{5}$$

and the number of ANs is:

$$K_{l^*} = 1 + (\Lambda_{\max} - 1) \sum_{i=0}^{l^*-1} (\Lambda_{\max} - 1)^i (P_{\max} - 1)^i. \tag{6}$$

The proof proceeds constructively. First, we prove that removing an edge connected to a PN does not change the

tree-structure form of a sub-graph with high probability. We, then, prove an equivalent result for an AN. Then, considering that a sub-graph with $l = 0$ trivially has a tree-structure, we prove a lower bound for the sub-graph of length $2l^*$.

Let us consider that $l < l^*$. Further, let us assume that $\mathcal{N}_{\vec{e}}^{2l}$ is tree-like and that k edges have been removed so far. The probability that removing another edge connected to a PN will not create a loop can be computed by considering whether expanding the sub-graph from that edge will not, at any level of the sub-tree, randomly hit an AN that is already in the neighborhood. This probability is equal to $\frac{(K-K_l-k)P_{\max}}{KP_{\max}-K_l-k}$. Assuming that K is sufficiently large, we find: $\frac{(K-K_l-k)P_{\max}}{KP_{\max}-K_l-k} = 1 - \frac{(K_l+k)(P_{\max}-1)}{KP_{\max}-K_l-k} \geq 1 - \frac{K_l^*}{K}$

Since $(K_{l+1} - K_l)$ ANs are added to the sub-graph at this step, the probability that the edge removal will lead to a tree-like sub-graph is $(1 - \frac{K_l^*}{K})^{(K_{l+1}-K_l)}$.

Equivalently for an AN, the probability that removing an edge connected to it does not create a loop is $\frac{(M-M_l-k)\Lambda_{\max}}{M\Lambda_{\max}-M_l-k}$. Assuming that M is sufficiently large, we find that:

$$\frac{(M - M_l - k)\Lambda_{\max}}{M\Lambda_{\max} - M_l - k} = 1 - \frac{(M_l + k)(\Lambda_{\max} - 1)}{M\Lambda_{\max} - M_l - k} \geq 1 - \frac{M_l^*}{M}.$$

Since $(M_{l+1} - M_l)$ PNs are added to the sub-graph at this step, the probability that this edge removal will lead to a tree-like sub-graph is at least: $(1 - \frac{M_l^*}{M})^{(M_{l+1}-M_l)}$.

We now transfer these results to the original sub-graph $\mathcal{N}_{\vec{e}}^{2l^*}$, where the probability that the sub-graph is tree-like is lower-bounded by

$$\mathbb{P}(\mathcal{N}_e^{2l} \text{is tree-like}) \geq \left(1 - \frac{K_{l^*}}{K}\right)^{K_{l^*}} \left(1 - \frac{M_{l^*}}{M}\right)^{M_{l^*}}.$$

We then use the Taylor series of $(1-x/n)^x$ and approximate the preceding bound with the first term of the series:

$$\mathbb{P}(\mathcal{N}_e^{2l} \text{ is tree-like}) \geq \left(1 - \frac{M_{l^*}^2}{M}\right)\left(1 - \frac{K_{l^*}^2}{K}\right).$$

This leads to

$$\mathbb{P}(\mathcal{N}_e^{2l} \text{ is not tree-like}) \leq \frac{M_{l^*}^2 + \frac{P_{\max}}{\Lambda_{\max}} K_{l^*}^2}{M}. \quad (7)$$

□

In order for the bound proposed by the above Lemma to correspond to a probability that converges to 0 as M grows, we need to make sure that $\gamma < M$. Based on the proof, and in particular Eq. (7), this suggests that the number of players needs to satisfy the following constraint:

$$M \geq 4 \cdot M_l^2 \quad (8)$$

The intuition behind this constraint is that the maximum number of pulls of a single player, D needs to be adjusted based on the total number of players, in order to ensure that the bipartite graph is sparse enough to ensure the resolution of collisions.

B. Analysis of the CR mechanism

In this section we establish the conditions under which the CR mechanism succeeds for all players. A CR round is successful for a player when they have found a collision-free arm within this round. Exiting a round without having found a collision-free arm is considered a failure for a player.

We begin by assuming that all sub-graphs are tree-like and then proceed with relaxing this assumption. The analysis consists of the following steps: first, we derive a condition under which the expected value of the probability of failure is monotonically decreasing with each iteration t of the CR round (Lemma B.1). Under this condition, which we refer to as the *stability condition* of the CR mechanism, a collision-free arm is found by all players with a probability that approaches 1 at a rate exponential in t (Lemma B.2). Then, we prove that the probability of failure concentrates around its expected value at a rate exponential in M, where the expectation is taken over all possible realizations of bipartite graphs (Lemma B.3). Finally, we show in Lemma B.3 that with high probability, an exponentially small number of players have not found a collision-free arm by a certain iteration t.

We consider a setting with K arms where the mean availability μ_k of arm k is randomly sampled in $[\mu_{\min}, 1]$. We assume that M players are using the CR mechanism described in Section 4. The following Lemma presents the stability condition of the CR mechanism. It is based on Lemma 1 of Luby et al. (1998), which derived a similar stability condition for a problem setting where arms corresponded to time slots instead of channels and were always available ($\mu_k = 1,\ \forall k \in \mathcal{K}$). Note that, in the following Lemma, t refers to an iteration within a single CR round.

Lemma B.1. *Consider a cycle-free bipartite graph derived by the edge-perspective degree distribution $\lambda(x)$. Denote with q_t the probability that a player has not found a collision-free arm at iteration t. Then, the probability that a player has not found a collision-free arm approaches 0 as t grows to infinity if, for all $q_t \in (0, 1]$:*

$$\lambda(1 - \rho(1 - q_t)) < q_t \quad (9)$$

Proof. We will first prove that the quantity on the left-hand side represents the probability of failure at the next iteration of the CR round, q_{t+1}.

Consider a PN of degree l. Denote by q the probability that an edge has not been removed, given that each of the other

$l-1$ edges has been removed with probability $1-p$. The edge of a player is removed when at least one of the other edges is removed. Thus, $q = p^{l-1}$. As edge-perspective degrees follow the degree distribution $\lambda(x)$, we can infer that

$$q_t = \sum_{l=1}^{D-1} \lambda_l p_{t-1} = \lambda(p_{t-1})$$

Similarly, consider an AN of degree l, where p denotes the probability that an edge has not been removed given that each of the other $l-1$ edges have been removed with probability $1-q$. As we know, the edge of an arm is removed when all other edges have been removed (no collisions) and the arm is available, which happens with probability μ_k, a random variable taking values in $[\mu_{\min}, 1]$. As a player keeps pulling the same set of arms until a reward of one is observed, we can ignore the effect of μ_k at this step and set $1-p = (1-q)^{l-1}$. Considering that the edge-perspective degrees of ANs follow the degree distribution $P(x)$, we can infer that

$$p_t = \sum_{l=1}^{D-1} p_l(1-(1-q_t)) = 1-\rho(1-q_t).$$

By inserting the expression of p_t into the expression of q_t, we get:

$$q_{t+1} = \lambda(1-\rho(1-q_t)).$$

In order for the CR mechanism to succeed we need to ensure that q goes to 0 as t grows. A necessary condition for this to happen is that $q_{t+1} < q_t,\ \forall q_t \in [0,1]$. The following expression is the stability condition of the CR mechanism:

$$\lambda\big(1-\rho(1-x)\big) < x, \forall x \in [0,1].$$

An alternative formulation of the stability condition that will prove useful in the analysis that follows is:

$$\lambda\big(1-\rho(1-x)\big) < x(1-\epsilon), \quad \forall x \in (0,1], \tag{10}$$

where ϵ is a positive constant.

□

Thus, in order to make sure that the CR mechanism succeeds with a target probability δ_5, we need to set $I_{\max}$ to the minimum number of time steps that satisfy $q_t < \delta_5$, multiplied by $1/\mu_{\min}$. This multiplication is due to the fact that any arm needs to be sampled at least $1/\mu_{\min}$ to return a reward of 1, which is necessary for the CR mechanism to continue.

The following trivial lemma, originally proposed by Luby et al. (1998), states that the probability of failure for a single player decreases exponentially with the iteration index t and that, for any upper bound on the probability of failure, there exists an iteration that satisfies it.

Lemma B.2. *If the stability condition in (10) is satisfied, then, for any $\gamma > 0$ we can set t to a constant such that $y_t < \gamma$.*

Proof. From the stability condition in Eq. (10) it holds that $x_t < x_{t-1}(1-\epsilon) < x_{t-2}(1-\epsilon)^2 < \cdots < (1-\epsilon)^t$. If we set $t = c/\epsilon$, for some $c > 1$, then $x_t < (1-\epsilon)^{c/\epsilon} \leq e^{-c}$, where the last inequality can be confirmed by studying the monotonicity of $\ln(1-x)+x$. We set $\gamma = e^{-c}$ and the proof is complete. □

Our analysis has so far assumed that all sub-graphs have a tree structure and does not take into account how performance on arbitrary graphs concentrates around its expected value. Using Lemma A.1, we can prove that all sub-graphs are tree-like with high probability. We, therefore, need to just study the concentration of the performance of tree-like sub-graphs around their expected value, denoted as q_t in Lemma B.1. We should note that the following Lemma, originally formulated by Luby et al. (2001, Theorem 1), is valid independently of whether the stability condition is satisfied.

Lemma B.3. *Let t denote the iteration in a CR round and Z_t be the random variable describing the fraction of players that have not found a collision-free arm after t iterations. Let $\mathbb{E}[Z_t]$ denote the expected value of Z_t, where the expectation is over all bipartite graphs and notice that it is equal to q_t, appearing in Lemma B.1. Then, there is a sufficiently large constant M, such that for any $\alpha > 0$ and some constant η:*

$$\mathbb{P}(|MZ_t - Mq_t| > M\alpha] < e^{-\eta\alpha^2 M} \tag{11}$$

Proof. The proof requires two intermediate steps. First, we need to bound the probability that the CR mechanism will create sub-graphs that do not have a tree structure. Then, we need to prove that the probability of failure for all graphs with a tree-structure concentrates around its expected value. We prove both results by formulating the edge removal process under the CR mechanism as a martingale and employing Azuma's inequality to prove a concentration bound.

Let M^* be the number of players for which the sub-graph of up to $2l$ levels is a tree. From Lemma A.1, we know that the probability that this sub-graph fails to be a tree is upper bounded by γ/M. For large M, this bound can be upper bounded as follows: $\gamma/M < \alpha/4$. Thus, the expected number of players with tree-structured sub-graphs is lower bounded as: $\mathbb{E}[M^*] \geq M(1-\alpha/4)$.

We now obtain a concentration result for M^* by describing edge removal as a martingale. We define Z_t to be the expected value for M^*, given the effect of the first t removals. In particular, $Y_0 = \mathbb{E}[M^*], Y_M = M^*$ and we define a filtration $\{\mathcal{F}_0, \cdots, \mathcal{F}_t\}$, where $\mathcal{F}_t$ is a σ-algebra

containing the sub-graph at step t. Then, the sequence $Z_t = \mathbb{E}[Y|\mathcal{F}_t]$ forms a standard Doob's martingale with $\mathbb{E}[Y_{t+1|Y_t}] = \mathbb{E}[Y_t]$. Using the additional observation that consecutive values of Y_t differ only by a constant (Luby et al., 2001, Lemma 1) and Azuma's inequality, we can derive the following concentration inequality for the number of players with sub-graphs without a tree-structure:

$$\mathbb{P}(|M^* - M| > M\alpha/2) < \frac{1}{e^{\eta_1 \alpha^2 M}}, \tag{12}$$

where η_1 is an appropriate constant.

Now, let M' denote the number of players, out of M^* total players, which have not found a collision-free arm after t steps. By definition, $\mathbb{E}[M'] = M^* q_t$. Again, we define Y_t as the expected value of M', given the results of the first t rounds. Since resolving collisions for a PN can only affect players in its sub-graph, the expression $|Y_{t+1} - Y_t|$ is a constant. Thus, using Azuma's inequality for the martingale $Z_t = \mathbb{E}[Y|\mathcal{F}_t]$ we get the following concentration result:

$$\mathbb{P}(|M' - M^* q_k| > M\epsilon/2) < \frac{1}{e^{\eta_2 \epsilon^2 M}}, \tag{13}$$

where η_2 is another constant. It is easy to verify that the random variables M, M^*, M' satisfy the following inequalities:

$$M' \leq M Z_l \leq M' + |M^* - M|, \tag{14}$$

where the final inequality is due to the observation that the inability of a player to resolve its collisions may be either due to that player not having a tree-structured sub-graph or having a tree-structured sub-graph but not having resolved a collision yet.

By combining the concentration inequalities in (12) and (13), we get a new concentration inequality:

$$\mathbb{P}(|M^* - M + M - M^* q_l| > M\epsilon) < \frac{1}{e^{\eta \epsilon^2 M}}$$
$$\rightarrow \mathbb{P}(|M Z_l - M^* q_l| > M\epsilon) < \frac{1}{e^{\eta \epsilon^2 M}},$$

where the last inequality is due to (14) and $\eta = \eta_1 + \eta_2$. We use the value $\eta = 1/(544 \bar{\Lambda}^{2l-1} \bar{P}^{2l})$ for this constant, as proposed by Richardson and Urbanke (2001, Theorem 2), who advised however that it does not lead to a tight bound. □

A direct conclusion from Lemma B.3 is that the probability that more than $\gamma' M$ players have not found a collision-free arm at iteration t is exponentially small in M. As M_t corresponds to the size of the sub-graph at time-step t, its value increases quickly with t and depends on $\bar{\Lambda}$. Thus, the right-hand side of (11) can be very large for small values of M. We should note that Lemma B.3 is only valid when the condition $M > 2\gamma/\alpha$ is satisfied.

C. Pseudocode for DYN-CR-UCB

We present the pseudocode of DYN-CR-UCB in Algorithm 2. In addition to the *preferences* list that a player updates when high quality arms are detected (Lines 27-29) and checks to find an arm to exploit (Lines 6-8), a player also updates an *unresolved* list, with arms that have given only zero consecutive rewards and are potentially being exploited by other players. If an arm remains in this list for more than L_2 time steps, it is transferred to the *occupied* list.

Algorithm 2 DYN-CR-UCB

1: Initialize $p = 0$, *occupied* = [], *preferences*= [], *unresolved* = [], phase="explore", L2 as in (4)
2: **while** phase =="explore" **do**
3: free, k = CR Round($\mathcal{K}, \Lambda, I_{\max}$)
4: $\hat{\mu}_k = S_k/T_k$
5: $B_t = 2\sqrt{\frac{\log T_m}{t}}$
6: **if** $k ==$ *Preferences*$[p]$ and *free* **then**
7: phase = "exploit"
8: **end if**
9: **if** *Preferences*$[p] \in$ *Occupied* **then**
10: $p = p + 1$
11: **end if**
12: **if** not *free* **then**
13: **if** k not in *unresolved* **then**
14: Insert k to *unresolved* {arm is potentially occupied}
15: $C_k = 1$
16: **else**
17: $C_k = C_k + 1$
18: **if** $C_k > L_2$ **then**
19: Insert k to *occupied* {arm is certainly occupied}
20: **end if**
21: **end if**
22: **else**
23: **if** k in unresolved **then**
24: $C_k = 0$ {arm is certainly not occupied}
25: **end if**
26: **end if**
27: **if** $\exists i,\ \mu_{min}[i] > \mu_{\max}[k] \forall k$ not in
28: *Preferences* and *Occupied* **then**
29: Insert k to *Preferences*
30: **end if**
31: **if** $\exists i$ not in *Preferences*[1:p] such that $\mu_{min}[i] > \mu_{\max}$[Preferences$[p]$] **then**
32: Insert *Preferences*$[p]$ to *Occupied*
33: **end if**
34: **end while**
35: Pull k until T^m {Exploitation phase}