Optimizing Media Access Strategy for Competing Cognitive Radio Networks

Youngjune Gwon
Harvard University
gyj@eecs.harvard.edu

Siamak Dastangoo
MIT Lincoln Laboratory
sia@ll.mit.edu

H. T. Kung
Harvard University
kung@harvard.edu

***Abstract*—This paper describes an adaptation of cognitive radio technology for tactical wireless networking. We introduce Competing Cognitive Radio Network (CCRN) featuring both communicator and jamming cognitive radio nodes that strategize in taking actions on an open spectrum under the presence of adversarial threats. We present the problem in the Multi-armed Bandit (MAB) framework and develop the optimal media access strategy consisting of mixed communicator and jammer actions in a Bayesian setting for Thompson sampling based on extreme value theory. Empirical results are promising that the proposed strategy seems to outperform Lai & Robbins and UCB, some of the most important MAB algorithms known to date.**

I. Introduction

Cognitive radios enable a new means to utilize spectrum, the scarcest resource in building wireless services. The fundamental premise of cognitive radio is an intelligent mechanism that identifies a vacancy in the spectral usage through sensing and learning, which can be implemented flexibly on programmable hardware. The majority of cognitive radio research has focused on dynamic spectrum access (DSA) [1], a compelling commercial model to improve the utility of a licensed spectrum and provide coexistence between the primary and secondary users of the spectrum. The secondary users are granted an *opportunistic* access as long as they can detect the primary users and relinquish the spectrum.

In this paper, we envision the use of cognitive radio technology for tactical wireless networking whose adverse operating environment includes malicious jammers and other security threats. We introduce the notion of Competing Cognitive Radio Network (CCRN), where a network of communicator (comm) nodes and jammers attempts to dominate the access to an open spectrum against a hostile opponent (another CCRN). In particular, we examine Multi-armed Bandit (MAB) problems [2], [3] to develop optimal CCRN channel accessing schemes.

The main contributions of this paper are two-fold. First, we provide an analytical framework for *competing* networks that can leverage the capability to jam their opponent by jointly coordinating with comm activities of own. The past approaches have been limited to an antijamming defense strategy for minimizing the adversarial attacks. Secondly, we have devised an optimal Bayesian setting for Thompson sampling, an old heuristic, to address the exploration-exploitation dilemma for the CCRN nodes taking actions (comm or jam) on a block of multi-channel spectrum and empirically validated its superior performance. The conjugate prior under an extreme-valued likelihood leads to superior performance over some of the most important MAB algorithms applied to our CCRN problem. The proposed method is also simple to implement.

The rest of this paper is organized as follows. In Section II, we explain our system model and underlying assumptions. Section III provides a brief background on MAB and presents mathematical formulation for CCRN. In Section IV, we examine three classes of algorithms for the stochastic MAB problems. We propose a new algorithm based on extreme value theory, conjugate prior, and Thompson sampling. Section V comparatively evaluates the existing and proposed MAB algorithms in a CCRN scenario featuring two tactical mobile networks. In Section VI, we provide the historical context of MAB, and Section VII concludes the paper.

II. Models and Assumptions

A. Competing Cognitive Radio Networks

For clarity of discussion, we assume *two* cognitive radio networks, namely *Ally* and *Enemy* CCRNs. Each CCRN consists of two types of cognitive radio nodes, communicator (comm) and jammer. The competitions are: 1) CCRNs try to achieve higher data throughput by adapting their comm transmissions to the opponent's jamming actions; and 2) CCRNs try to decrease the opponent's data throughput by jamming each other's comm activities. To devise a winning media access strategy, we need to jointly optimize antijamming and jamming schemes.

B. Networking Model

Mobile ad hoc network (MANET) best describes the networking model discussed in this paper. The primary-secondary user dichotomy used in DSA is no longer valid, and limited or no fixed infrastructural support for the nodes is assumed. A CCRN can adopt a centralized control model where the node actions—*i.e.*, which channels should a comm node transmit data or should a jammer jam—are computed by a singular decision maker to assure the coherent, network-wide strategy. On the other hand, a distributed control model allows each node to compute its own action. We consider both control models in this paper.

This work is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

C. Communication Model

Fig. 1 depicts a time-slotted, multi-channel spectrum for open access. There are N non-overlapping channels located at the center frequency f_i (MHz) with bandwidth B_i (Hz) for $i = 1, \ldots, N$. Each time-frequency slot, represented by a tuple $\langle f_i, B_i, t \rangle$, gives a transmission (Tx) opportunity and has a duration of T msec. We assume that a node can sense other nodes' transmissions in range. Such sensing capability, however, is not coupled to specifics of any conventional media access control mechanisms such as CSMA.

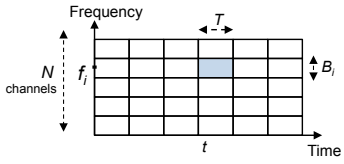

Fig. 1. Tx opportunities in slotted multi-channel spectrum for open access

D. Reward Model

The reward metric evaluates the performance of a CCRN. A comm node receives a reward of B (bits) upon a successful transmission, which requires only one comm node transmitting in the given slot without being jammed. If there were two or more simultaneous comm transmissions (from either the same or different network), a collision occurs, and no comm node gets a reward. A successful jamming results in the same reward value for the opposing comm node's transmission attempt that would have been successful otherwise. For example, an Ally jammer receives B when it jams an Enemy comm node transmitting B-bit worth of data. If there were no jamming, the Enemy comm would have earned B. Also, it is possible for an Ally jammer to jam an Ally comm mistakenly (*e.g.*, due to faulty intra-network coordination), which we call *misjamming*.

III. Multi-armed Bandit Formulation for Competing Cognitive Radio Networks

Thompson [2] introduced Multi-armed Bandit (MAB) to address the problem in the clinical trial of a medical treatment causing different effects to patients. This section presents the MAB formulation for CCRNs with the goal of accumulating optimal rewards from unknown parameters of the channel-node interactions that need to be learned sequentially.

A. Notation and Preliminaries

We make a simplifying assumption that each CCRN has C comm nodes and J jammers (typically, $C + J < N$). We use superscripted ^{t} for 'at time t,' not the 'power of.' The Ally and Enemy node actions at time t are $a_A^t = \{a_{A,comm}^t, a_{A,jam}^t\}$ and $a_E^t = \{a_{E,comm}^t, a_{E,jam}^t\}$ containing both comm and jamming actions, the size-C vectors $a_{A,comm}^t$, $a_{E,comm}^t$ and the size-J $a_{A,jam}^t$, $a_{E,jam}^t$. An ith element in $a_{A,comm}^t$ designates the channel number that the ith Ally comm node tries to access at t. Similarly, a jth element in $a_{A,jam}^t$ is the channel that the jth Ally jammer tries to jam at t.

Ω^t, whose element designates each channel's state (an integer value), is a size-N vector that describes the outcome of the Ally and Enemy node actions used to determine the reward at time t:

$$a_A^t \times a_E^t \longrightarrow \Omega^t$$

It is more convenient to compute a reward from each channel (than node), and we use $r_{A,k}^t$ to designate the instantaneous reward for Ally resulted from channel k at time t. The total reward at t is the sum over all N channels: $R_A^t = \sum_{k=1}^{N} r_{A,k}^t$.

For illustrative purposes, let $C = 2$, $J = 2$, $N = 10$. If $a_{A,comm}^t = [7\ 3]$, Ally comm node 1 transmits in channel 7, and Ally comm node 2 in channel 3 at time t. $a_{A,jam}^t = [1\ 5]$ means that Ally jammers jam channels 1 and 5 at t. Let $a_{E,comm}^t = [3\ 5]$ and $a_{E,jam}^t = [10\ 9]$. Fig. 2 depicts the resulting channel-action bitmap where 1 indicates transmit or jam and 0 otherwise. Ally jammer 2 on channel 5 is successful whereas jammer 1 is not. The comm transmissions collide in channel 3, but Ally has a successful comm transmission in channel 7. Thus, the Ally comm 1 and jammer 2 receive a reward of B each. Enemy has no success in comm or jamming.

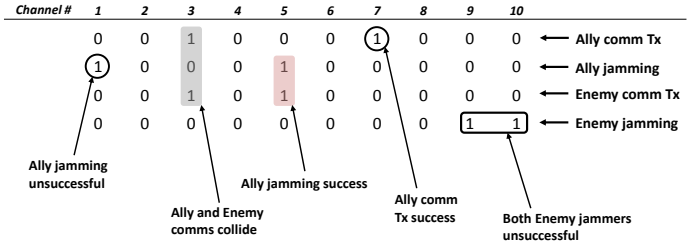

Fig. 2. Channel-action bitmap example

The Ally network strategy σ_A^t is a function over time. It takes necessary information such as sensing results and past action-outcome/reward statistics as input and determines Ally node actions. Under the centralized decision making, we write:

$$\{x_A^j\}_{j=1}^t, \{a_A^j, \Omega^j\}_{j=1}^{t-1} \xrightarrow{\sigma_A^t} a_A^t$$

where x_A^t is the Ally sensing results at t.

Under the distributed decision making, each node in the network computes its own action. For node i in Ally (whether it is a comm node or jammer):

$$x_{A,i}^t, \{x_A^j, a_A^j, \Omega^j\}_{j=1}^{t-1} \xrightarrow{\sigma_{A,i}^t} a_{A,i}^t$$

where $x_{A,i}^t$ is the sensing information only available to node i at time t, and $\sigma_{A,i}^t$ the strategy of node i's own. At time t, node i does not yet have all sensing results except its own $x_{A,i}^t$. For the distributed case, node strategies can differ, and there is no guarantee that conflicting actions of the nodes in the same network such as collision and misjamming are resolved.

B. Multi-armed Bandit (MAB)

MAB is best explained with a gambler facing N slot machines (arms). The gambler's objective is to find a strategy that maximizes $R^t = \sum_{j=1}^{t} r^j$ for some t, the *cumulative*

reward over a finite horizon. Lai & Robbins [3] introduced the concept of *regret* for a strategy σ measuring the distance from optimality

$$\Gamma^t = t\mu^* - \mathbb{E}\left[R_\sigma^t\right]$$

where μ^* is the hypothetical, maximum average reward if gambler's action were resulting the best possible outcome each round, and R_σ^t the actual reward achieved with σ. Γ^t is mathematically convenient, and maximizing the *expectation* of R^t turns out to be equivalent to minimizing Γ^t.

Lai & Robbins [3] further derived the mathematical qualification for an optimal strategy:

$$\lim_{t\to\infty} \sup \mathbb{E}[T_i^t] \leq \frac{\log t}{\mathrm{D}_{KL}(p_i \parallel p^*)} \tag{1}$$

where sup means supremum, T_i^t is the total number for arm i being played, and $\mathrm{D}_{KL}(. \parallel .)$ is the Kullback-Leibler divergence [4] measuring the dissimilarity between the probability distributions p_i and p^* for the i-th arm's reward and the maximum reward resulted by choosing only the best possible arm each time. Eq. (1) provides the least upper bound for the number of times should an optimal arm—which could be different each time—be played asymptotically. Lai & Robbins also provided an algorithm that satisfies the condition of Eq. (1), which will be discussed in Section IV.

C. MAB Model for Competing Cognitive Radio Network

We can now explain the MAB model for CCRN. An arm corresponds to a channel in the spectrum under competition. Comm nodes and jammers are the players that the networks allocate to play (*i.e.*, transmit or jam) the channels. Since each network has multiple nodes, our problem is classified as multi-player MAB [5], which is different from the classic single-player MAB formulated by Lai & Robbins [3]. In addition, we have two system variations depending on whether a centralized control entity or each player makes the play decisions.

Fig. 3 illustrates the CCRN with a central decision maker (*e.g.*, base station, super node) computing the network-wide strategy and disseminating *all* node actions. It is assumed for the centralized multi-player MAB that the decision maker should be able to collect sensing results from each player and the exact outcome of every play in order to make sound decisions over time.

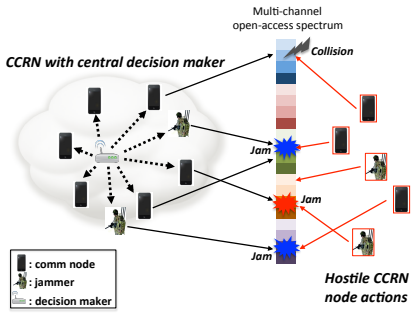

Fig. 3. Centralized multi-player MAB for CCRN

Fig. 4 illustrates the case for distributed decision making. Here, each node makes its own play decision based on the information collected in best effort compared to the centralized multi-player MAB that requires the tight intra-network communication to collect information and disseminate the strategy. After each play, the node observes the outcome, computes its reward, and maintains its play statistics that can be shared with others in the network.

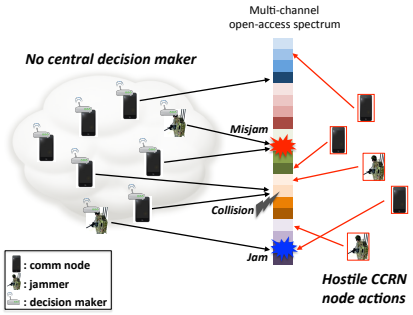

Fig. 4. Distributed multi-player MAB for CCRN

D. Problem Statement

The MAB formulation for CCRN makes us consider the problem of sequentially sampling the total network reward from the N channel reward populations $r_1^t, r_2^t, \ldots, r_N^t$ over time. The rewards are manifested by the mixed player actions from the same and opposing networks that dynamically affect the outcome each time. Differentiated from the classic MAB problems, the player action in CCRN comprises an action (transmit) and its *anti*-action (jam). The anti-action does not draw the reward directly from a channel but can deprive that generated by a comm node. Formally, we search for an optimal strategy σ_{opt}^t that minimizes the growth of regret Γ^t:

$$\sigma_{opt}^t = \arg\min_\sigma \Gamma^t = \min_\sigma \{\mathbb{E}[\sum_{i=1}^{M}\sum_{j=1}^{t} r_{(i)}^j] - \mathbb{E}\left[R_\sigma^t\right]\} \tag{2}$$

To express the regret in Eq. (2), we use $r_{(i)}^t$ an *ordered* sequence of the N instantaneous channel rewards at time t such that $r_{(1)}^t \geq r_{(2)}^t \geq \ldots \geq r_{(N)}^t$. There are $M = C + J$ total number of nodes in a network, thus summing M highest rewarding channels reflect the optimal allocation of players.

IV. Formulation of Optimal Strategies

This section examines three different classes of MAB algorithms and proposes a new algorithm that outperforms the existing algorithms empirically. We present pseudo-code, assuming there is only one ($M = 1$) node in the network for simplicity of explanation. This simplification reduces the problem into correctly guessing the most rewarding channel over time.

A. Asymptotically Optimal Algorithm

Described in Algorithm 1, Lai & Robbins [3] keeps track of cumulative reward R_i^t and total number of accesses T_i^t for channel i, and draws two candidate channels c_{MPE} and c_{RR}, based on the maximum point estimate (MPE) criterion (*e.g.*, channel with highest sample mean) and round robin (RR) selection, respectively. The Kullback-Leibler divergence between the two serves a test statistic to finalize the choice.

Algorithm 1 (Lai & Robbins)

while $t < 1$ ▷ initialized offline
  Access each channel at least once
  Record $R_i^t = \sum_{j=1}^{t} r_i^j$ and T_i^t for every channel i
end
while $t \geq 1$ ▷ online
  Compute $\mu_i = R_i^t / T_i^t \ \forall i$
  Find MPE candidate $c_{MPE} = i^*$ s.t. $\mu_{i^*} = \max \mu_i$
  Find RR candidate $c_{RR} = (t \mod N) + 1$
  if $\mathrm{D}_{KL}(p_{RR} \parallel p_{MPE}) > \log(t-1)/T_{c_{RR}}^t$
    Access c_{MPE} and observe $r_{c_{MPE}}^t$
    Update $R_{c_{MPE}}^t$ and $T_{c_{MPE}}^t$
  else
    Access c_{RR} and observe $r_{c_{RR}}^t$
    Update $R_{c_{RR}}^t$ and $T_{c_{RR}}^t$
  end
end

The essence of Lai & Robbins is to consider exploitation (choosing the MPE candidate) vs. exploration (choosing the arbitrary RR candidate). The condition $\mathrm{D}_{KL}(p_{RR} \parallel p_{MPE}) > \log(t-1)/T_{c_{RR}}^t$ assures that choosing the MPE candidate is optimal after a sufficient number of exploratory trials.

B. Indexing

Despite its algorithmic simplicity, Lai & Robbins comes down to estimating D_{KL} accurately, which is computationally difficult from sampling. The second class of MAB algorithms uses indexing that is computable substitute for D_{KL}. Auer *et al.* [6] formulated an indexing scheme called Upper Confidence Bound (UCB) presented in Algorithm 2.

Algorithm 2 (UCB)

while $t < 1$ ▷ initialized offline
  Same as that of **Algorithm 1**
end
while $t \geq 1$ ▷ online
  Compute point estimate $\mu_i = R_i^t / T_i^t \ \forall i$
  Compute index $g_i = \mu_i + \sqrt{\alpha \log \frac{t}{T_i^t}} \ \forall i$
  Access channel $i^* = \arg\max_i g_i$
  Update $R_{i^*}^t$ and $T_{i^*}^t$
end

C. Thompson Sampling

The last class of algorithms uses a probability matching technique known as Thompson sampling [2] that selects actions according to their probability of being optimal. It is largely a heuristic and has reemerged in the recent machine learning literature such as Agrawal & Goyal [7], which provides the most rigorous mathematical treatment available to date. The full proof of Thompson sampling on its convergence, however, remains to be an open problem. It is best understood under a Bayesian setup as in Algorithm 3.

Algorithm 3 (Thompson Sampling)

Require: $d = \{x, a, r\}$ for context x, action a, reward r, estimator $p(\theta|d) \propto p(r|x,a,\theta)p(\theta)$ parameterized by θ
while $t \geq 1$ ▷ online
  Acquire x^t
  Draw $\theta^t \sim p(\theta)$
  Choose a^t to access $i^* = \arg\max_i \mathbb{E}[r_i^t | x^t, \theta^t]$
  Observe actual r^t
  Update $d = d \cup \{x^t, a^t, r^t\}$
  Update $p(\theta) = p(\theta|d)$
end

D. Our Algorithm

We propose a new MAB algorithm based on extreme value theory [8], conjugate priors, and Thompson sampling.

1) Distribution of maximum reward sequence: Let $Y^t = \max\{r_1^t, \ldots, r_N^t\}$ where r_i^t represents the reward from channel i at t. Since the sequence $Y^1, Y^2, \ldots, Y^t$ consists only of the maximum channel reward each time, it must have achieved the distribution p^* in Eq. (1). Furthermore, the sequence should result in an *upper bound* of the optimal mean reward μ^*. Therefore, all we need is our strategy σ to empirically follow the distribution of Y^t. But how is it distributed?

Fisher & Tippet [9] and Gnedenko [10] proved the existence of limiting distributions for block maxima (or minima) of random variables. Their findings became the foundation of extreme value theory used widely in financial economics.

Theorem 1: (Fisher & Tippett, Gnedenko) Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables and $M_n = \max\{X_1, \ldots, X_n\}$. If real number pairs (a_n, b_n) exist such that $a_n > 0$ and $lim_{n\to\infty} P(\frac{M_n - b_n}{a_n} \leq x) = F(x)$, where $F(.)$ is a non-degenerate distribution function, then the limiting distribution $F(.)$ belongs to only Fréchet, Gumbel, or Weibull family of probability distribution functions.

Proof: See Fisher & Tippet [9] and Gnedenko [10]. ■

2) Conjugate priors: In Bayesian inference, the posterior is updated by the observed likelihood given the prior distribution:

$$\underbrace{p(\theta|r)}_{\text{posterior}} \propto \underbrace{p(r|\theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}$$

When the probabilistic model for the likelihood is known, we can set the prior and posterior distributions conveniently of the *same* family of functions. This is known as conjugate prior. Since the reward distribution under our search is extreme-valued, our likelihood choices are left to Fréchet, Gumbel, or Weibull distributions. Table I summarizes the conjugate priors having an extreme valued likelihood distribution [11].

TABLE I
BAYESIAN CONJUGACY UNDER EXTREME-VALUED LIKELIHOOD

Distribution family	Likelihood model	Conjugate priors
Fréchet	Pareto	Gamma
	Lognormal	Gamma or normal
Gumbel	Exponential	Gamma
	Normal	Normal
	Gamma	Gamma
Weibull	Weibull	Inverse gamma
	Beta	Unknown

3) The algorithm: In summary, our algorithm presented in Algorithm 4 performs Thompson sampling that follows an extreme-valued likelihood and updates the posterior distribution based on its conjugate prior. However, we need to decide on which extreme value distribution is suitable for CCRN.

Since both Fréchet and Gumbel distributions model *unbounded* random variables, we adopt a *Weibull* likelihood with the inverse gamma conjugate prior (see Table I), reasoning that the maximum reward value for competing mobile networks should be *finite*. The lack of theoretical analysis on Thompson sampling makes it difficult to justify our design choice. In Section V, we show an empirical evidence that backs up our choice. However, the search for the best choice remains open at least until we explore all possibilities listed in Table I.

A Weibull distribution has finite endpoints. Its conjugate prior, the inverse gamma distribution, has two hyperparameters $a, b > 0$. Our algorithm draws the scale parameter θ from the inverse gamma prior $p(\theta|a,b) = \frac{b^{a-1}e^{-b/\theta}}{\Gamma(a-1)\theta^a}$ for $\theta > 0$ where a and b are the sample mean and variance of the reward of a channel. The Weibull random variable generated by θ drawn from the prior estimates the expected reward for the channel. After observing the actual reward, the posterior update follows.

Algorithm 4 (Proposed Algorithm)

Require: $a_i, b_i = 0\ \forall i$
1: **while** $t < 1$ ▷ initialized offline
2: Access each channel until $a_i, b_i \neq 0\ \forall i$, where a_i and b_i are sample reward mean and variance
3: **end**
4: **while** $t \geq 1$ ▷ online
5: Draw $\theta_i \sim$ inv-gamma(a_i,b_i)
6: Estimate $\hat{r}_i =$ weibull(θ_i,β_i) $\forall i$ for given $0.5 \leq \beta_i \leq 1$
7: Access channel $i^* = \arg\max_i \hat{r}_i$
8: Observe actual $r^t_{i^*}$ to update $\{R^t_{i^*}, T^t_{i^*}\}$
9: Update $a_{i^*} = a_{i^*} + T^t_{i^*}$, $b_{i^*} = b_{i^*} + \sum_t (r^t_{i^*})^{\beta_{i^*}}$
10: **end**

V. EMPIRICAL EVALUATION

A. Evaluation Scenarios and Metric

We evaluated the centralized and distributed multi-player MAB scenarios for the two CCRNs, Ally and Enemy, in a custom MATLAB simulator. Ally network ran the MAB algorithms explained in Section IV while Enemy network was configured with the baseline *static* and uniformly *random* strategies. We assumed that the central decision maker had perfect knowledge (*i.e.*, all nodes' sensing results) in the decision making for the centralized scenario. On the contrary, every node in the distributed scenario was a decision maker, using its own sensing results only (no information sharing).

We adopted the average reward per channel as the performance evaluation metric for a CCRN:

$$\bar{R}^t = \frac{1}{N \cdot t} \sum_{j=1}^{t} \sum_{i=1}^{N} r_i^j$$

where r_i is the ith channel reward, and there are N channels in the spectrum. We used the following channel reward model:

- $r_i^t = 1$ if only one comm node transmits and no jamming in channel i at time t;
- $r_i^t = 1$ if a jammer jams the opposing comm node's transmission;
- $r_i^t = 0$ otherwise (*e.g.*, collision, jamming).

B. Tested Algorithms

We tested a multi-player version of Lai & Robbins (L&R-M) [5], multi-player UCB (UCB-M) [6], and our algorithm against the baseline static and random strategies. In static strategy, nodes initially choose to access some channels and continue to access the same channels throughout. Random strategy chooses a uniformly random channel for each play. Additionally, we tested Z-heuristic from Rivest and Yin (Z) [12]. Like our algorithm, Z-heuristic is based on Thompson sampling but uses a Gaussian likelihood. Conjugate prior for the Gaussian likelihood is also Gaussian.

For our algorithm and Z-heuristic, we use the following multi-player technique. Select the best channel from estimating $\hat{r}_i$. Remove the selected channel and regenerate $\hat{r}_i$ for the *remaining* $N-1$ channels. Select the best among the remaining. We repeat the process until we allocate all M nodes to play.

C. Results and Discussion

The simulated spectrum has $N = 10$ channels. For each CCRN, we vary the number of comm nodes $C = 2, 4, 8$, but fix the number of jammers to $J = 2$. Comm nodes have a transmit probability $p_{tx} = 0.5$ whereas jammers jam with probability 1. We run $t = 1,000$ iterations and measure steady-state, cumulative average rewards for comparison.

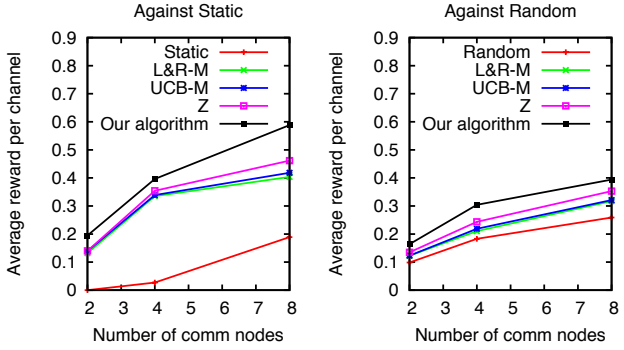

Fig. 5. Performance of tested algorithms in centralized scenario

Fig. 5 compares the performance of tested algorithms in the centralized scenario. We plot the reward performance of our algorithm (Algorithm 4), L&R-M, UCB-M, and Z-heuristic tested against the baseline strategies (static and random). The baseline performances plotted are against our algorithm. We can clearly observe performance advantage of our algorithm over L&R-M, UCB-M, and Z-heuristic. The proposed algorithm can learn static transmission and jamming patterns effectively. Static strategy yields near-zero reward at $C = 2$. Since we have fixed $J = 2$, static strategy can realize nonzero rewards when $C > 2$. Learning is harder against random strategy because randomization gives an effective exploration mechanism. Random strategy, however, can only explore. This lack of exploitation explains its poor performance overall.

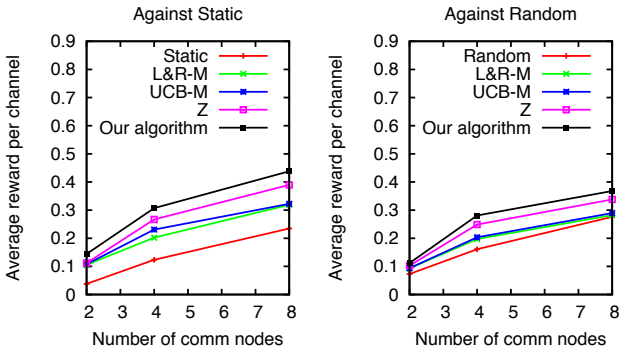

Fig. 6. Performance comparison in distributed scenario

In Fig. 6, we compare the performance in the distributed scenario. Again, the proposed algorithm achieves the best performance. The lack of explicit intra-network cooperation among the nodes seems to be critical, costing more than 15% degradation in the performance. This loss is the result of more collisions and misjamming in the Ally CCRN. Consequently, we observe a slight performance gain for the Enemy CCRN.

VI. Related History of MAB Problems

In 1933, Thompson [2] introduced a stochastic MAB problem and proposed an optimal heuristic known as Thompson sampling, which remains to be an effective action selection strategy that often outperforms modern proposals. Robbins 1952 [13] presented the first sequential analysis of the single-player MAB problem. In Bellman 1954 [14], MAB problems were formulated as a class of Markov decision process (MDP). Gittins 1979 [15] proved the existence of a Bayes optimal indexing scheme for MAB problems if they can be modeled as a stationary MDP. Lai & Robbins 1985 [3] introduced the notion of regret, derived its lower bound using the Kullback-Leibler divergence, and constructed asymptotically optimal allocation rules. Anantharam *et al.* 1987 [5] extended Lai & Robbins for multi-player. Whittle 1988 [16] introduced PSPACE-hard restless MAB problems and showed that suboptimal indexing schemes are possible. Rivest & Yin 1994 [12] proposed Z-heuristic that achieved a better empirical performance than Lai & Robbins. Auer *et al.* 2002 [6] proposed Upper Confidence Bound (UCB), an optimistic indexing scheme.

VII. Conclusions and Future Work

We have described competing cognitive radio networks (CCRNs) that operate under hostile assumptions to strive for dominating access to an open spectrum. Our notion of CCRN advocates both communications and jamming capabilities. We have adopted the MAB framework and thoroughly examined classical solutions known to date to develop a novel, optimal media access strategy for CCRN.

An optimal CCRN strategy should embrace randomized algorithms although doing randomization only will lead to poor performance because a strategy needs to exploit its learning. Our results indicate that Thompson sampling proves to be most effective in addressing the exploration-exploitation tradeoff, which is fundamental to construct a MAB-optimal strategy for CCRN. Our proposed algorithm is only slightly more complex than Thompson sampling, but could consistently outperform the existing MAB algorithms. For our next step, we plan to study application scenarios with the outcome of actions depending on unknown geographic environments that govern radio propagation behavior, more complex reward models, and parameter optimization. In addition, we want to address the issues in sensing errors and failover. Protocol specification and implementation on software radios are also on the way.

References

[1] Q. Zhao and B. Sadler, "A Survey of Dynamic Spectrum Access," *IEEE Signal Processing Magazine*, May 2007.

[2] W. R. Thompson, "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[3] T. L. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.

[5] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for Multiarmed Bandit Problem with Multiple Plays–Part I: I.I.D. Rewards," *IEEE Trans. on Automatic Control*, vol. 32, no. 11, pp. 968–976, Nov 1987.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, May 2002.

[7] S. Agrawal and N. Goyal, "Analysis of Thompson Sampling for the Multi-armed Bandit Problem," in *Proc. of 25th Annual Conference on Learning Theory (COLT)*, 2012.

[8] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction*. Springer, 2006.

[9] R. A. Fisher and L. H. C. Tippett, "Limiting Forms of the Frequency Distribution of the Largest and Smallest Member of a Sample," *Proc. Cambridge Phil. Soc.*, pp. 180–190, 1928.

[10] B. V. Gnedenko, "Sur la distribution limite du terme maximum d'une serie aleatoire," *Annals of Mathematics*, pp. 423–453, 1943.

[11] E. George, U. Makov, and A. Smith, "Conjugate Likelihood Distributions," *Scandinavian Journal of Statistics*, pp. 147–156, 1993.

[12] R. L. Rivest and Y. Yin, "Simulation Results for a New Two-armed Bandit Heuristic," in *Workshop on Computational Learning Theory and Natural Learning Systems*, 1994.

[13] H. Robbins, "Some Aspects of the Sequential Design of Experiments," *Bulletin of American Mathematics Society*, vol. 58, pp. 527–535, 1952.

[14] R. Bellman, *A Problem in the Sequential Design of Experiments*. Defense Technical Information Center, 1954.

[15] J. C. Gittins, "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 148–177, 1979.

[16] P. Whittle, "Restless Bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25A, pp. 287–298, 1988.