# Learning to coordinate without communication in multi-user multi-armed bandit problems

**Orly Avner** ORLYKA@TX.TECHNION.AC.IL

**Shie Mannor** SHIE@EE.TECHNION.AC.IL
*Department of Electrical Engineering*
*Technion - Israel Institute of Technology*
*Haifa, Israel*

**Editor:**

## Abstract

We consider a setting where multiple users share multiple channels modeled as a multi-user multi-armed bandit (MAB) problem. The characteristics of each channel are initially unknown and may differ between the users. Each user can choose between the channels, but her success depends on the particular channel as well as on the selections of other users: if two users select the same channel their messages collide and none of them manages to send any data. Our setting is fully distributed, so there is no central control and every user only observes the channel she currently uses. As in many communication systems such as cognitive radio networks, the users cannot communicate among themselves so coordination must be achieved without direct communication. We develop algorithms for learning a stable configuration for the multiple user MAB problem. We further offer both convergence guarantees and experiments inspired by real communication networks.

**Keywords:** Bandits,Multi-user,Stable marriage.

## 1. Introduction

The inspiration for this paper comes from the world of distributed multi-user communication networks, such as cognitive radio networks. These networks consist of a set of communication channels with different characteristics, and independent users whose goal is to transmit over these channels as efficiently as possible. Such a setup introduces several challenges. First, the networks' distributed nature prohibits any form of central control. In addition, many users operate on an "ad hoc" basis, preventing them from forming inter-user communication. In fact, they probably do not know how many users share their network.

On top of these issues, that concern the coordination of multiple users, the channel characteristics may be initially unknown, and differ between users. The first assumption entails a learning approach, while the consequences of the different characteristics are discussed in detail in the sequel.

### 1.1 Cognitive radio networks

Cognitive Radio Networks (CRNs), introduced in Mitola and Maguire (1999), have attracted considerable attention in recent years. The idea that lies at the heart of CRNs is that advanced sensing mechanisms and increased computation power may enable radio devices to dramatically improve their performance in terms of resource utilization, resilience and more. In our paper we focus on developing a sensing and transmission scheme that enables users to learn a stable, orthogonal configuration without communicating directly.

### 1.2 Multi-armed bandits

A well known framework for learning in CRNs is the classical Multi-Armed Bandit (MAB) model. MABs offer a simple framework for learning the characteristics of a number of unknown options in an online manner, while balancing exploration and exploitation. A MAB problem consists of a single user repeatedly choosing between arms with different, initially unknown, characteristics. After every round, the user acquires a reward that depends on the arm she chose. Her goal in most setups is to maximize the expected sum of rewards acquired over time.

The channels of a CRN are naturally cast as the arms of a bandit, as first suggested in Jouini et al. (2010), with different performance measures (bandwidth, ack signals, bit rate) serving as the reward.

Many papers propose solutions for the stochastic MAB problem (see, e.g., Auer et al. (2002a); Garivier and Cappé (2011)) and its adversarial version (see, e.g., Auer et al. (2002b)), but they all assume a single user is sampling the arms of the bandit. However, this assumption does not apply in multi-user networks. In the multi-user MAB model, users compete over the arms of *the same* bandit. As a result, they are bound to experience collisions (i.e., multiple users sampling the same arm), unless they employ some form of collision avoidance or coordination mechanism. Collisions in communication networks result in performance degradation, corresponding to reward loss in the MAB model. In order to meet the goal of reward maximization, the presence of multiple users must be addressed. We survey several approaches to this issue in Section 1.4.

### 1.3 Extension of the CRN-MAB setting

The novelty introduced in our paper lies in the combination of bandit learning, multiple users, different reward distributions for different users and no direct communication. The combination of these last two demands - different distributions and no direct communication, poses a real challenge.

As explained in detail in Section 2.3 and in Section 2.4, the only thing we can guarantee in terms of network behavior in this setup is stability, which we define formally in Definition 1. In a dynamic, distributed network, stability should not to be taken for granted, and it is of great value. Once a network has reached a stable configuration, users can focus on utilizing its resources, rather than engaging in coordination or learning efforts; a stable network is more robust and efficient. Reaching stability is a nontrivial task, since users must learn their channel characteristics while coordinating their actions with the other users, based on very limited observations.

### 1.4 Previous work

We now present several approaches to our problem, coming from different areas and disciplines. Our problem may be viewed as an assignment problem, i.e., maximum weight matching in a weighted bipartite graph. Several papers have been published on the distributed assignment problem, but to the best of our knowledge none of them offers a solution for our problem. The well-known Hungarian method requires full knowledge of the graph (i.e., channel characteristics) and assumes the existence of central control. The Bertsekas auction algorithm of Bertsekas (1988) frees us from the need for central control, at the cost of direct communication between nodes. The classical Gale-Shapley algorithm solves the problem of finding a stable marriage configuration, assuming the weights are known. Some papers like Cohen et al. (2013); Leshem et al. (2012) have actually applied it to CRNs, but not in the learning context. Another noteworthy work in this context is Amira et al. (2010), which addresses the challenge of limiting communication between nodes to a minimum, but does not consider learning. Two additional results that deal with distributed stable marriage offer lower bounds and state that *some* form of information exchange is inevitable when solving such problems are Gonczarowski and Nisan (2014); Kipnis and Patt-Shamir (2009).

The papers closest to ours in spirit are those dealing with multi-user MABs.Anandkumar et al. (2011) and Avner and Mannor (2014) consider reward distributions that do not vary between users. The latter introduces an algorithm that is able to cope with a variable number of users. Another paper, that does address different reward distributions for different users, is Kalathil et al. (2012). The authors employ the Bertsekas auction algorithm, enabling users to reach a reward-maximizing solution, at the price of direct communication.

We would like to point out that communication between users is undesirable not only because of its price in terms of network resources and time. Once users depend on communication, they are vulnerable to intentional attacks disrupting it, as well as noise bursts that are common in CRNs. They also need to have some knowledge about each other in order to set up a communication protocol - knowledge that our algorithm does not require.

## 2. Model and formulation

In this section we describe the model, the assumptions accompanying it and our goal.

### 2.1 System and users

We model a communication network with $K$ channels, servicing $N$ independent users. Our work is based on the assumption that $K \geq N$, which is reasonable since without it, implementing a time division based mechanism is necessary. Once such a mechanism is applied, the assumption that $K \geq N$ is valid again. Time is slotted and users' clocks are synchronized, also a mild assumption for modern communication systems.

The communication network consists of $K$ channels, where only one user can transmit over a certain channel during a single time slot. Each transmission yields a reward, which we assume to be stochastic.

The users are a group of $N$ independent, selfish agents. Their observations are local, consisting only of the history of their actions and rewards. In addition, they do not know the number of users they share a network with. There is no central control managing their use of the network, and they do not have direct communication with each other.

A key characteristic of our model is that the expected reward a channel yields depends not only on the identity of the channel, but also on the identity of the user. Formally, the rewards of the channels are Bernoulli random variables with expected values $\{\mu_{n,k}\}$, where $n \in \{1, \ldots, N\}$ and $k \in \{1, \ldots, K\}$.

We model the users' sharing resources by representing the network using a *single* bandit. Two users attempting to access the same channel at the same time will experience a collision. The result of a collision is loss of communication for that time slot for the colliding users, i.e., zero reward. A user $n$ that accesses a channel $k$ alone during a certain time slot will receive a reward drawn i.i.d. from a Bernoulli distribution with expected value $\mu_{n,k}$. Throughout the paper, we use the term *configuration* to refer to a mapping of users to channels.

### 2.2 Limited coordination

In an effort to keep our model faithful to real world CRNs, we limit the coordination between users to a minimum. Thus, users can only transmit in a channel of their choice, or sense the spectrum range and receive binary feedback regarding all channels $\{1, \ldots, K\}$ at time $t$. A "0" represents no transmission in channel, while "1" stands for the opposite.

### 2.3 Reward maximizing solution

We adopt a system-wide view for characterizing the optimal solution. The goal is to maximize the sum of rewards over all users, over time. The solution is an orthogonal configuration, in which each user uses a single channel, and channels are sampled by no more than one user. The assignment of users to channels is chosen so as to maximize the sum of rewards: $R^* = \max_{\pi \in \mathcal{C}} \sum_{n=1}^{N} \mu_{n,\pi(n)}$, where $\mathcal{C}$ is the set of all possible permutations of subsets of size $N$ chosen without replacement from the set $\{1, \ldots, K\}$.

However, reaching such a solution requires direct information exchange. Assume channel $k$ is optimal for two different users $m$ and $n$, but $\mu_{m,k} > \mu_{n,k}$. To maximize the system-wide reward, user $n$ must step down and choose a different channel. The lack of central control requires explicit information exchange regarding the values of $\mu_{m,k}$ and $\mu_{n,k}$, for $m$ and $n$ to decide which of them should step down.

Since the reward-maximizing solution is not attainable in our setup, due to the limited information exchange, we focus on convergence to a stable, orthogonal configuration.

### 2.4 Stable marriage solution

Our goal is to develop policies that will lead users to a stable configuration. We employ the notion of stable marriage to formally define stability:

**Definition 1** *A Stable Marriage Configuration (SMC) is an assignment of users to channels such that no two users would be willing to swap channels, had they known the true values of the expected rewards. Formally, for a pair of users n,m:*

$$S_1 \triangleq (\mu_{n,a_n} < \mu_{n,a_m}) \qquad \textit{user n would like to swap}$$

$$S_2 \triangleq (\mu_{m,a_m} \leq \mu_{m,a_n}) \qquad \textit{user m is willing like to swap,}$$

*where $a_m$ and $a_n$ are the users' current actions. In an SMC,*

$$S_1 \wedge S_2 = 0 \quad \forall n, m.$$

### 2.5 Goal

Given a system with $K$ channels and $N$ users, allowing only limited communication as described in Section 2.2, our goal is to reach a configuration that is orthogonal: no two users use the same channel, and an SMC, according to Definition 1.

## 3. Coordination protocol

Our coordination protocol balances the limitations of Section 2.2 with the users' need for information exchange by introducing a signalling mechanism between pairs of users. At predefined time slots, a user wishing to occupy a channel may transmit in that channel to express her wish. In order to ensure that this signal is received by the user currently occupying the channel, we employ a frame-based protocol. We assume users can transmit and receive at the same time, a reasonable requirement in modern communication systems.

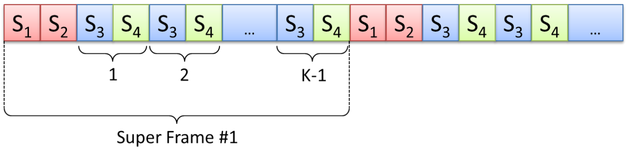


Figure 1: Coordination protocol - frame structure.

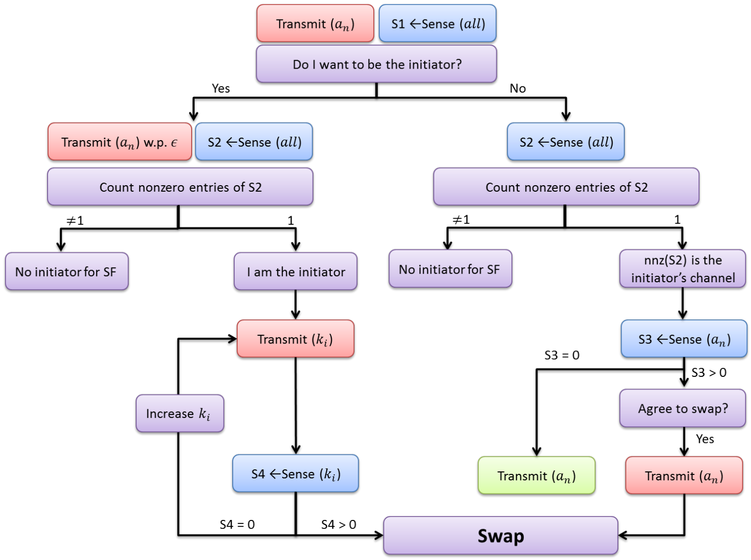


Figure 2: Coordination protocol flow for a single super frame. Events (transmission or sensing) appearing on the same line take place simultaneously.

The following explanation is best understood by observing Figure 1 and Figure 2. Our protocol divides time into super frames of length $T_{\text{SF}} = 2 + 2\,(K-1) = 2K$. Each super frame begins with a pair of time slots, $S_1$ and $S_2$, during which a single initiator is coordinated for the entire super frame, as described in Algorithm 1. These are followed by $K-1$ mini-frames of 2 time slots each, denoted by $S_3$ and $S_4$. Each of these mini-frames corresponds to one channel on the initiator's list of preferred channels. Thus, a single super frame enables one user to go over her entire preference list and signal other users, suggesting they swap channels with her. Time slots marked $S_4$ allow users who are not occupied with coordination to sample their current channel and proceed with the learning process. Thus, all but two users (initiator and responder) gather a sample during each mini-frame, resulting in at least $K-2$ samples per super-frame for each of the users, except for the initiator.

While this may seem like much coordination, the protocol is very simple to implement, and is indeed lightweight when compared to other protocols. The following lemma quantifies the time devoted to signalling.

**Lemma 1** *In every super-frame, $(K-1)(N-2)$ learning samples are gathered by all users combined. During this period, $4K$ signalling and sensing actions are performed by all users combined, so the ratio between signalling and learning is*

$$L \triangleq \frac{4K}{(K-1)(N-2)}.$$

The ratio $L$ can be improved if the number of users is known. Also, it clearly improves as the number of users grows.

## 4. The CSM-MAB algorithm

We propose a user-level algorithm for a fully distributed system. When all users in the network apply CSM-MAB, described in Algorithm 1, the assignment of users to channels is guaranteed to be orthogonal, and converges to an SMC.

Our algorithm begins with a start up phase, during which users transmit and sense in order to reach an initial orthogonal configuration (line 1). This phase follows the lines of the CFL algorithm introduced in Leith et al. (2012), and converges quickly.

At the beginning of each SF, users execute the **rank_channels** procedure to individually create a list of channels they prefer over their current action (line 4). Channels are assigned values according to their UCB indices, calculated using the well known formula from Auer et al. (2002a) : $I_{n,k}(t) = \hat{\mu}_{n,k} + \sqrt{\frac{2\ln t}{s_{n,k}}}$, where $\hat{\mu}_{n,k}$ is the empirical mean of the reward acquired by user $n$ on channel $k$ up till time $t$ and $s_{n,k}$ is the number of times she sampled arm $k$ up till time $t$.

Next, the users coordinate an initiator: every user who would like to improve upon her current channel presents herself as the initiator with a probability of $\epsilon = \frac{1}{K}$ (lines 5-11). An agreed initiator for the current SF emerges if and only if the number of non-zero entries in $S_2$ is exactly 1 (the value of $\epsilon$ is chosen in order to maximize the probability of this occurring). Once a single initiator is agreed upon, all users take note of her current channel, based on $S_1$. They will need this knowledge to decide whether to accept her swapping suggestion.

The initiator proceeds to signal other users, based on her ranking of channels (lines 13-21). Signalling is implemented in **propose_swap** by transmitting in the initiator's channel

**Algorithm 1** CSM-MAB algorithm

```
$a_n(0) \leftarrow$ **apply_CFL**$(K)$
**for all** frames $t$ **do**
  **if** $\mod(t, T_{\text{SF}}) == 1$ **then** {Beginning of SF}
    $list \leftarrow$ **rank_channels**$(a_n(t-1), \hat{\mu}_n, s_n)$
    **if** $list \neq \mathbf{0}$ **then** {User $n$ would like to change channels}
      $flag_n \leftarrow$ **rand**(Bernoulli, $\epsilon$)
      **if** $flag_n == 1 \wedge flag_i == 0 \ \forall i \neq n$ **then** {Only $n$ raised a flag}
        $initiator = n$ {User $n$ is the initiator for this SF}
        $pref = 1$ {Swapping preference is initialized to 1}
      **end if**
    **end if**
  **else**
    **if** $(initiator == n) \wedge (pref > 0)$ **then** {$n$ is the initiator, $list$ not exhausted yet}
      $response \leftarrow$ **propose_swap**$(list(pref))$
      **if** $response == 1$ **then** {Responder agreed or channel is available}
        $a(t) \leftarrow$ **swap**$(a_n(t), list(pref))$
        $pref \leftarrow 0$
      **else**
        $pref \leftarrow pref + 1$ {Advance to next best channel}
      **end if**
    **end if**
  **end if**
  $r_n(t) \leftarrow$**execute_action**$(a_n(t))$
  **update_stats**$\big(r_n(t), \hat{\mu}_{n,a_n(t)}, s_{n,a_n(t)}\big)$
**end for**
```

**note:** $\hat{\mu}_{n,k}$ is the empirical mean of the reward for user $n$ on channel $k$;
$s_{n,k}$ is the number of times user $n$ has sampled arm $k$.

of interest (time slot $S_3$ in Figure 2). Each responder (i.e., signalled user) checks whether swapping channels with the initiator will improve her situation, based on her own ranking. Once a responder agrees, a swap takes place. No more signalling attempts are made till the end of the SF, and users simply continue sampling their chosen channels. If the responder refuses, the initiator will approach the next-best channel on her list. She will continue the process until she (a) finds a partner that agrees to swap; or (b) exhausts her list of potential swaps.

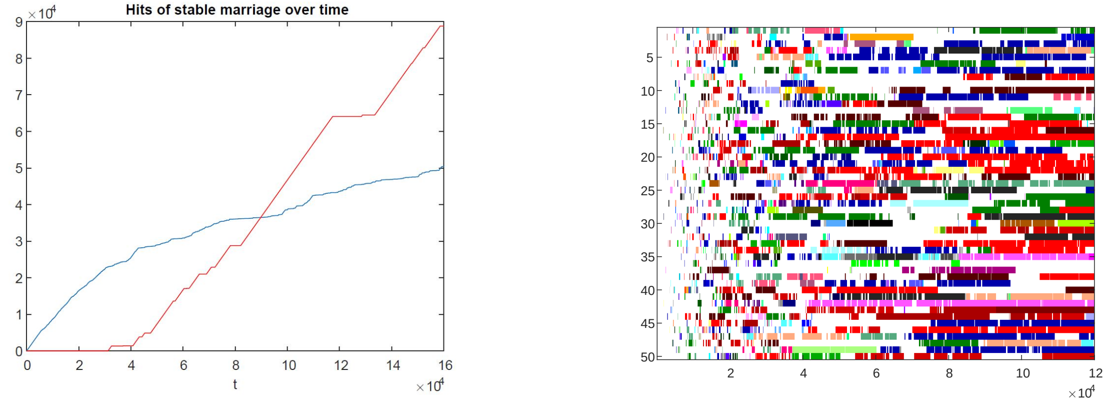
If the initiator would like to switch to a vacant channel, then there is no need for signalling, and she simply updates her chosen action. All users except for the initiator and the responder gather a sample for the learning process with each mini-frame (lines 23-24).

Our main theoretical result is stated in Theorem 1.

**Theorem 1** *Consider a system with $K$ channels and $N$ users, with channel rewards characterized by the matrix $\boldsymbol{\mu}$. Applying CSM-MAB (Algorithm 1) by all users will result in convergence to an orthogonal SMC: For all $\delta > 0$ there exists $T(\delta)$ such that for all time slots $t > T$, the probability of the system's being in an SMC is at least $1 - \delta$.*

The proof of Theorem 1 consists of two aspects: orthogonality and stability. Orthogonality is a rather direct result of the definition of Algorithm 1 and the initialization phase, while stability is a more involved result, proved by defining a system-wide potential function. Both appear in the extended version of the paper, Avner and Mannor (2015).

## 5. Experiments



(a) Total time spent in SMCs - classical UCB (blue) vs. CSM-MAB (red).

(b) Convergence to SMC: horizontal axis shows time, lines are realizations. Colored pixels represent stable configurations.

Figure 3: Comparison to classical UCB and convergence to SMC

In order to demonstrate the merits of our algorithm, we implemented a simulation of a distributed multi-user communication network. In this network, users cannot communicate with each other directly. However, they can listen to all channels and transmit over a channel of their choice, updating this choice with each time slot.

A user transmitting over a channel receives a binary reward, drawn i.i.d. from a Bernoulli distribution with parameter $\mu_{n,k}$. We present results obtained with $K = 12$ channels and $N = 10$ users transmitting for $T = 120000$ time slots, averaged over 50 repetitions.

First, we compare the performance of our coordinated approach to an approach that applies a learning algorithm without addressing the presence of multiple users, by examining convergence to a stable state. Figure 3a shows the cumulative number of times our policy "hits" an SMC, together with the number of times a classic UCB policy does so. Our algorithm achieves this often, while the classic UCB struggles, especially as time advances.

Next, we examine convergence to different SMCs over several repetitions of a certain setup. Figure 3b shows that the periods of time users spend in unstable configurations decrease as the experiment advances, and that users move between different SMCs.

The number of times users change their choice of channel is also of interest when stability is the goal. Figure 4 shows the cumulative number of changes per user, over time. Different users have different patterns, depending on the difficulty of their problem: users that have small differences between channel characteristics will experience more policy changes.

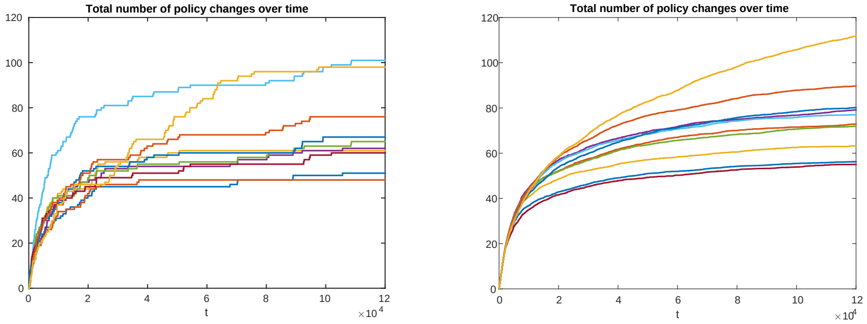


Figure 4: Number of changes in users' choice of channels. On the left - a single realization. On the right, the number of changes, averaged over 50 repetitions.

## 6. Conclusion

We have presented an extension of the multi-user MAB problem for the case of different reward distributions between users, together with limited information exchange. Using a specialized signalling method, our algorithm enables multiple users to learn network characteristics and converge to an orthogonal configuration that is also a stable marriage. We also provide a theoretical performance guarantee. Finally, we present the results of an experimental setup and examine different aspects of our approach's performance.

In the future we intend to extend our work to a dynamic scenario, both in terms of channel characteristics and number of users. The latter should be straightforward due to the minimal inter-dependency of users.

## References

N. Amira, R. Giladi, and Z. Lotker. Distributed weighted stable marriage problem. In *Structural Information and Communication Complexity*, pages 29–40. Springer, 2010.

A. Anandkumar, N. Michael, A.K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *Selected Areas in Communications, IEEE Journal on*, 29(4):731–745, 2011.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

O. Avner and S. Mannor. Concurrent bandits and cognitive radio networks. In *European Conference on Machine Learning*, 2014.

O. Avner and S. Mannor. Learning to coordinate without communication in multi-user multi-armed bandit problems. 2015. URL `http://arxiv.org/abs/1504.08167`.

D.P. Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of operations research*, 14(1):105–123, 1988.

K. Cohen, A. Leshem, and E. Zehavi. Game theoretic aspects of the multi-channel aloha protocol in cognitive radio networks. *Selected Areas in Communications, IEEE Journal on*, 31(11):2276–2288, 2013.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference On Learning Theory*, pages 359–376, 2011.

Y.A. Gonczarowski and N. Nisan. A stable marriage requires communication. *arXiv preprint arXiv:1405.7709*, 2014.

W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *Signals, Circuits and Systems (SCS), 2009 3rd International Conference on*, pages 1–6. IEEE, 2010.

D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multi-player multi-armed bandits. In *51st IEEE Conference on Decision and Control*, 2012.

A. Kipnis and B. Patt-Shamir. A note on distributed stable matching. In *IEEE International Conference on Distributed Computing Systems*, 2009.

D.J. Leith, P. Clifford, V. Badarla, and D. Malone. WLAN channel selection without communication. *Computer Networks*, 2012.

A. Leshem, E. Zehavi, and Y. Yaffe. Multichannel opportunistic carrier sensing for stable channel access control in cognitive radio systems. *Selected Areas in Communications, IEEE Journal on*, 30(1):82–95, 2012.

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