AN EFFICIENT POLICY FOR D2D COMMUNICATIONS AND ENERGY HARVESTING IN COGNITIVE RADIOS: GO BAYESIAN!

S. J. Darak, Honggang Zhang†, Jacques Palicot‡, Christophe Moy‡*

*ECE, Indraprastha Institute of Information Technology, Okhla Phase III, New Delhi 110020, India
† ISEE, Zhejiang University, Zheda Road 38, Hangzhou 310027, China
‡ SCEE, CentraleSupélec/IETR, CS 47601, 35576 Cesson-Sévigné Cedex, France

ABSTRACT

Recently, there has been a surge of interests in paradigms such as device-to-device (D2D) communications and radio frequency energy harvesting (RFEH) to improve the spectrum as well as energy efficiencies of next-generation decentralized cognitive radio networks. However, little attention has been paid to the dual but competing task of subband selection of any desired bandwidth in D2D mode (i.e., opportunistic vacant spectrum access) and RFEH mode as well as need to minimize the subband switching cost (SSC) for an efficient implementation. Taking these factors into account, a new D2D-RFEH policy is proposed. It consists of: 1) Bayesian approach based Tunable Thompson Sampling (TTS) algorithm to learn subband statistics, 2) Subband access scheme employing TTS algorithm for minimizing collisions among the secondary users, and 3) Mode selection scheme. The simulation results, complexity and SSC analysis validate the superiority of the proposed policy over the policies employing frequentist approach based learning algorithms.

Index Terms— Device to device communications, Radio frequency energy harvesting, Thompson Sampling

1. INTRODUCTION

Further boosting the spectrum and energy efficiencies are the primary goals of next-generation wireless communication networks such as 5G and Long Term Evolution-Advanced (LTE-A) [1–5]. As a result, there has been a surge of interests in the promising techniques such as device-to-device (D2D) communications and radio frequency energy harvesting (RFEH) from the academia and industrial partners [1–5]. The D2D communications in cognitive radio networks (CRNs) allow direct communication between secondary users (SUs) over identified vacant licensed subband(s). The RFEH is a new trending technique which facilitates the conversion of received RF signals into electricity, in the range of mW, which can be supplied for data transmission. It is expected that more efficient techniques will be available in the near future along with an international standard for RFEH.

Dynamic spectrum learning and access (DSLA) in decentralized CRNs consisting of multiple SUs is an important research problem [6–9]. Nonetheless, this task has now become more challenging in CRNs consisting of RFEH enabled SUs. This is because, occupied subbands with higher RF energy are the optimum subbands in RFEH mode while the vacant subbands of any desired bandwidth and lower RF energy are the optimum subbands in the D2D mode. Furthermore, tunable bandwidth requirements of 5G and LTE-A must be taken into account to support wide variety of services ranging from data to multimedia. From energy efficiency perspective, subband switching cost (SSC), i.e. total penalty incurred in terms of reconfiguration delay, energy consumption and protocol overhead when a SU switches the frequency subband, should be as minimum as possible. The design of an efficient D2D-RFEH policy with tunable bandwidth access, lower SSC and computational complexity is the objective of the proposed work.

The proposed D2D-RFEH policy, for next-generation decentralized CRNs of multiple RFEH enabled SUs, consists of: 1) Tunable Thompson Sampling (TTS) online learning algorithm to learn subband statistics, 2) Subband access scheme for orthogonalization of SUs, and 3) Mode selection scheme to decide when to switch from D2D mode to RFEH mode and vice versa. To the best of our knowledge, the proposed D2D-RFEH policy is the first which is designed using Bayesian approach based TTS algorithm compared to existing policies which are designed using frequentist approach based learning algorithms. Additionally, the proposed subband access scheme employs TS algorithm to minimize SSC, as against the randomization approach in existing policies. The simulation results validate the superiority of the proposed policy over existing policies. In the next Section, system model and literature review are discussed. The proposed policy is presented in Section 3 followed by the performance analysis in Section 4. Section 5 concludes the paper.

2. SYSTEM MODEL AND LITERATURE REVIEW

Consider the slotted CRN consisting of multiple primary users and M RFEH enabled SUs. The desired vacant bandwidth of SUs is denoted by $B_v(k), k \in \{1, 2, ..., M\}$. Consider the bandwidth of wideband input signal is divided into N uniform subbands of smallest PU channel bandwidth, B_{cmin} where B_{cmin}=(1/N) on the normalized frequency scale. Let i^{th} subband switches its state from being vacant to occupied and vice versa according to discrete Markov process with probabilities, p^i_{vo} and p^i_{ov}, respectively. Then, using Markov chain analysis, steady-state probabilities of subband being vacant, denoted as $P_{vac}(i), i \in \{1, 2, ...N\}$, are [5]

$$P_{vac}(i) = \frac{p^i_{ov}}{p^i_{ov} + p^i_{vo}} \ \forall i \quad (1)$$

Basic assumptions are: 1) Infrastructure-less decentralized CRN where all SUs employ the same policy but do not exchange any information with other SUs, 2) D2D and RFEH modes use same hardware interface which means that simultaneous D2D communications and RFEH are not possible, 3) SU can sense only one subband in each time slot, 4) Subbands occupied by SUs are not useful for RFEH due to lower RF energy, 5) $P_{vac}(i)$, $\forall i$ are unknown to SUs.

In D2D mode, when multiple SUs transmit on the non-orthogonal frequency subbands, collision occurs and SU does not get any reward i.e., $B_c(k)$ =0. When no such collision occurs, it is assumed that the SU transmits successfully and gets the positive integer reward, $B_c(k) = (B_v(k)/B_{cmin}), \forall k$. Then, $r_u(k, t) = r_u(k, t-1) + B_c(k)$. For each data transmission attempted over bandwidth B_{cmin}, the battery charge is reduced by β_d. In RFEH mode, for each occupied subband of bandwidth, B_{cmin} and energy β, the battery capacity is increased by $\beta_s = \beta - \beta_r$ where β_r is charge for harvesting bandwidth, B_{cmin}. It can be safely assumed that $\beta_r = \xi \cdot \beta_d$, $\beta_d = \theta \cdot \beta_s$ where $\xi < 1$ and $\theta \geq 1$. Let $S^*(t)$ and $S(t)$ denote the total reward of the genie-aided policy (i.e., the policy where $P_{vac}(i)$, $\forall i$ are known a priori and no collisions among SUs) and the decentralized D2D-RFEH policy, respectively. Then, the total regret $U(t)$ of the CRN upto time t is given by Eq. 2 [6–9] and should be as small as possible.

$$U(t) = S^*(t) - S(t) = \sum_{k=1}^{M}\sum_{v=0}^{t-1}[r^*_u(k, v) - r_u(k, v)] \quad (2)$$

2.1. DSLA in Decentralized CRNs without RFEH

Various DSLA schemes have been proposed for orthogonalization of SUs to optimum vacant subbands in decentralized CRNs [6–9]. In policy, ρ^{rand} [6], each SU randomly and independently choses the rank, $R(k) \in \{1, 2, ..M\}$ in the beginning. In each time slots, underling learning algorithms such as upper confidence bound (UCB), $\varepsilon-$greedy, etc., calculate the quality index for each subband. Then, the SU with the rank $R(k)$ choses the subband with the $R(k)^{th}$ best quality index. Another policy in [7] follows time division fare share approach where the rank of each SU is rotated in circular fashion from 1 to M to allow an equal access to the optimum subbands among all SUs. Though both policies are mathematically proved to be optimal, the SSC of [7] is very high compared to that of [6]. In [8, 9], we proposed variable filtering architecture and its integration with tunable extension of ρ^{rand} , ρ^{t_rand}, to support the tunable bandwidth requirements of SUs. However, average SSC of [6–9] is high.

2.2. DSLA in Decentralized CRNs with RFEH

Recent research efforts investigating the different ways to adopt RFEH in CRNs are available in [2–5]. In [2], D2D-RFEH policy for centralized CRNs has been investigated while the case of decentralized CRNs is taken in this paper. The optimization based policy in [3] and learning based policy in [4] consider the subband selection for RFEH enabled SUs in CRNs while decision making policy in [5] deals with the switching actions between two modes with known subband statistics. Though all these policies are proved to be optimal, they are designed for single SU CRNs which means that the SU collisions and their effect on the total reward of the decentralized CRNs has not been considered. An optimization based policy in [3] is computationally complex and may not be feasible for battery-operated resource-constrained SU terminals. Hence, online learning algorithm based policies need to be investigated. Furthermore, the use of Bayesian approach based learning algorithms to minimize SSC without compromising on the reward/regret as well as the effect of mode switching decisions on the performance of D2D-RFEH policy have not been explored yet in the literature.

3. PROPOSED D2D-RFEH POLICY

Proposed policy comprised of three sub-sections which are:

3.1. Online Learning Algorithm

As discussed, the task of an online learning algorithm has become more challenging due to RFEH mode. A learning algorithm must identify vacant subbands of any desired bandwidth with low RF energy as well as occupied subband with higher RF energy. The design of such algorithm is discussed below.

The basic idea of Thompson Sampling (TS) algorithm is to assume some prior distribution on the probability statistics of each subband (e.g. uniform prior) and at any time slot, sample the subband according to its posterior probability of being the optimum [10, 11]. The main reasons for chosing TS algorithm for the proposed D2D-RFEH policy are:

1. An online learning algorithm is said to be optimal in subband selection task if [10–12]

$$\liminf_{t\to\infty} \frac{\mathbb{E}[\bar{T}(k, i, t)]}{\ln t} \geq \frac{1}{K(P_{vac}(i), P_{vac}(i^*))}, \ \forall i \quad (3)$$

where

$$i^* = \arg\max_i P_{vac}(i) \quad (4)$$

where $\bar{T}(k,i,t)$ denotes the number of time slots upto t slots where the subband i is sensed by k^{th} SU, $K(p,q)$ denotes the Kullback-Leibler divergence factor and equality sign in Eq. 3 corresponds to an asymptotically optimal algorithm. Though frequentist approach based UCB, optimization based KL-UCB [12], Bayesian approach based TS and Bayes-UCB algorithms [10] have been proved to be asymptotically optimal for Bernoulli rewards, it has been just recently proved that the scaling constant of Bayesian algorithms is better than that of others [10, 11].

2. Bayesian subband sampling approach allows these algorithms to have lower SSC than UCB algorithm or its extensions through empirical observations.
3. TS is a least complex Bayesian algorithm since it requires just one sample from the posterior of subband statistics when compared to optimization problem in KL-UCB and computation of quantiles in Bayes-UCB [10–12].

To the best of our knowledge, usefulness of TS algorithm for DSLA applications in decentralized CRNs has not been studied yet. Furthermore, existing TS algorithm needs to be re-designed to take into account the tunable bandwidth scenarios, i.e., $B_v(k) > B_{cmin}$. For example, in D2D mode, when $B_v(k) = B_{cmin}$, then the corresponding k^{th} SU can chose any one of $N_D(k) = N$ orthogonal subbands and hence, the SUs selecting distinct subbands would not experience collision. However, when $B_v(k) > B_{cmin}$, the number of subband choices for the corresponding k^{th} SU is reduced to $N_D(k) = [N + 1 - B_c(k)]$ non-orthogonal subbands. In such case, the SUs transmitting on distinct subbands may experience collision due to non-orthogonality of subbands. By taking these factors into account, TTS algorithm is proposed which is given in Algorithm 1.

Algorithm 1: Tunable Thompson Sampling Algorithm for k^{th} Secondary User in D2D Mode

Parameters: $N, M, N_D(k), B_v, R(k), l_k \in \{1,2,..,N_D(k)\}$
Input: $r_s(k,:,t-1), \mathrm{T}_D(k,:,t-1)$
Output: $I_D(k,t)$

If $(any(\mathrm{T}_D(k,:,t-1) == 0))$
$I_D(k,t) = l_k \; s.t. \mathrm{T}_D(k,l_k,t-1) = 0$
Else
1. $\alpha_k = \exp\left(\left\lceil \frac{\sum_{p=1}^{M} \lceil B_v(p)\cdot N \rceil}{B_v(k)\cdot N} \right\rceil \cdot B_v(k)\right), \quad \forall l_k$
2. $B(k,l_k) = Beta(\alpha_k \cdot \mathrm{r}_s(k,l_k,t-1)+1, \mathrm{T}_D(k,l_k,t-1) - \alpha_k \cdot \mathrm{r}_s(k,l_k,t-1)+1) \quad \forall l_k$
3. $I_D(k,t) = l_k$ *corresponding to* $R(k)^{th}$ *max. value of* $(B(k,l_k))$

End

In the proposed TTS algorithm, parameter, α_k, is introduced to minimize the number of collisions due to non-orthogonality of subbands and the rank, $R(k)$ is obtained using the scheme discussed in Section 3.2. If l_k^{th} subband (which corresponds to $I_D(k,t)$ in Algorithm 1) is sensed as vacant by k^{th} SU at time t, then

$$r_s(k,l_k,t) = r_s(k,l_k,t-1) + 1 \tag{5}$$

$$\bar{r}_s(k,i_k,t) = \bar{r}_s(k,i_k,t-1) + 1, \; \forall i_k \tag{6}$$

where $i_k \in \{l_k, l_k+1, ..., l_k - 1 + B_c(k)\}$. Otherwise, $r_s(k,l_k,t) = r_s(k,l_k,t-1)$ and $\bar{r}_s(k,i_k,t) = \bar{r}_s(k,i_k,t-1), \; \forall i_k$. On the other hand, if l_k^{th} subband is observed as occupied with its RF energy higher than $(\beta_d \cdot B_c(k)/\theta)$, then

$$r_{rf}(k,l_k,t) = r_{rf}(k,l_k,t-1) + 1 \tag{7}$$

$$\bar{r}_{rf}(k,i_k,t) = \bar{r}_{rf}(k,i_k,t-1) + 1, \; \forall i_k \tag{8}$$

Otherwise, $r_{rf}(k,l_k,t) = r_{rf}(k,l_k,t-1)$ and $\bar{r}_{rf}(k,i_k,t) = \bar{r}_{rf}(k,i_k,t-1), \; \forall i_k$. Finally,

$$T_D(k,l_k,t) = \sum_{p=1}^{t} 1_{\{I_D(k,p)=l_k\}} \tag{9}$$

$$\bar{T}(k,i_k,t) = \sum_{p=1}^{t} 1_{\{I_D(k,p)\supseteq i_k\}}, \; \forall i_k \tag{10}$$

$$X_D(k,l_k,t) = \frac{r_s(k,l_k,t)}{T_D(k,l_k,t)}, \; \bar{X}(k,i_k,t) = \frac{\bar{r}_s(k,i_k,t)}{\bar{T}(k,i_k,t)}, \; \forall i_k \tag{11}$$

$$Y_D(k,l_k,t) = \frac{r_{rf}(k,l_k,t)}{T_D(k,l_k,t)}, \; \bar{Y}(k,i_k,t) = \frac{\bar{r}_{rf}(k,i_k,t)}{\bar{T}(k,i_k,t)}, \; \forall i_k \tag{12}$$

where $X_D(k,l_k,t)$ (or $\bar{X}(k,i,t)$, resp.) denotes the learned probability of l_k^{th} (or i^{th}, resp.) subband being vacant at k^{th} SU, $Y_D(k,l_k,t)$ (or $\bar{Y}(k,i,t)$ resp.) denotes the learned probability of l_k^{th} (or i^{th}, resp.) subband having RF energy higher than $(\beta_d \cdot B_c(k)/\theta)$ (or (β_d/θ), resp.) at k^{th} SU.

In the RFEH modes, $R(k) = 1, \; \forall k$ and the task of the TTS algorithm is to identify contiguous set of occupied subbands with higher RF energy where the contiguous bandwidth, $B_o(k)$, should not be greater than but close to analog front-end bandwidth, B_{afe}. Note that $B_o(k)$ in RFEH mode is equivalent to $B_v(k)$ in D2D mode but $B_v(k)$ is chosen by SU while $B_o(k)$ is chosen by TTS algorithm. Here, only $B_{afe} = B_{cmin}$ case is considered for the brevity of the paper and the case of $B_{afe} > B_{cmin}$ will be considered in extension of this work. Hence, $N_{RF}(k) = N$ and $B_c(k) = 1$. The proposed TTS algorithm for RFEH mode is shown in Algorithm 2. Similar to D2D mode, r_s, $\bar{r}_s$, r_{rf} and $\bar{r}_{rf}$ are updated in each time slot based on sensing outcome of $I_{RF}(k,t)$. Similarly, T_{RF}, X_{RF}, Y_{RF} (which are equivalent to T_D, X_D, Y_D, respectively in D2D mode) are also updated. In this way, TTS algorithm learns from the sensing outcomes in each mode.

3.2. Subband Access Scheme

The task of an access scheme, when implemented at all SUs in decentralized CRNs, is to orthogonalize SUs operating in D2D mode to the optimum vacant subbands. Since there are no collisions among SUs in RFEH mode, orthogonalization is not required. The design of the access scheme involves

Algorithm 2: Tunable Thompson Sampling Algorithm for k^{th} Secondary User in RFEH Mode

Parameters: $N, N_{RF}(k), B_o, l_k \in \{1,2,..,N_{RF}(k)\}$
Input: $r_{rf}(k,:,t-1), T_{RF}(k,:,t-1)$
Output: $I_R(k,t)$

If $(any(T_{RF}(k,:,t-1) == 0))$
$I_{RF}(k,t) = l_k \ s.t.\ T_{RF}(k,l_k,t-1) = 0$
Else
1. $\alpha_k = \exp(B_o(k)), \quad \forall l_k$
2. $B(k,l_k) = Beta(\alpha_k \cdot r_{rf}(k,l_k,t-1)+1,$
$T_{RF}(k,l_k,t-1) - \alpha_k \cdot r_{rf}(k,l_k,t-1)+1) \quad \forall l_k$
3. $I_{RF}(k,t) = l_k$ *corresponding to* *max. value of* $(B(k,l_k))$

End

two sub–problems: 1) Selection of new rank, $R(k)$ when the corresponding k^{th} SU collides with other SUs, and 2) Tunable range of rank, $L_t(k)$ (referred to as tunable subset size in [8, 9]). The discussion in this section is restricted to (1) while solution to (2) has been proposed in [8,9].

Taking into account the superiority of learning algorithms based subband selection schemes [6–9] over the random subband selection scheme, a TS based rank assignment scheme, replacing the conventional random rank assignment scheme [6–9], is proposed as shown in Algorithm 3. Here, $T_R(k,r_k,t)$ denotes the number of time slots out of t where $R(k) = r_k,\ r_k \in \{1,2,..,L_t(k)\}$ and $C(k,r_k,t)$ denotes the number of time slots out of $T_R(k,r_k,t)$ where k^{th} SU experiences collision. Note than $R(k)$=0 when SU is in RFEH mode and $R(k) = I_R(k,t-1)$ in D2D mode. In [6,8,9], each SU randomly updates their rank, $R(k)$, after collision. This approach is not fair to SUs with $B_v(k) > B_{cmin}$ since the number of collisions, and hence, SSC of such SU is higher compared to SUs with narrower $B_v(k)$. The first if loop in the Algorithm 3 eliminates this unfairness. In the proposed scheme, the use of TS algorithm guarantees that all SU will eventually settle in different rank thereby leading to better scaling constant than randomization based rank assignment schemes [6–9] and hence, faster orthogonalization of SUs.

Algorithm 3: Thompson Sampling Algorithm for Rank Selection at k^{th} Secondary User in D2D Mode

Parameters: $L_t(k), B_o, r_k \in \{1,2,..,L_t(k)\}$
Input: $r_r(k,:,t-1), T(k,:,t-1)$
Output: $I_R(k,t)$
If $(C_R(k,R(k),t) \ mod \left(\frac{B_v(k)}{B_{cmin}}\right) == 0)$
If $(any(T_R(k,:,t-1) == 0))$
$I_R(k,t) = r_k \ s.t.\ T_R(k,r_k,t-1) = 0$
Else
1. $B(k,r_k) = Beta(r_r(k,r_k,t-1)+1,$
$T_R(k,r_k,t-1) - r_r(k,r_k,t-1)+1) \quad \forall r_k$
2. $I_R(k,t) = r_k$ *corresponding to* *max. value of* $(B(k,:))$

End
Else $I_R(k,t) = I_R(k,t-1)$
End

3.3. Mode Selection Scheme

In the proposed policy, each SU independently makes the mode switching decision based on the available battery charge, W_B ($0 \leq W_B \leq 1$) along with the given probabilities of switching to RFEH mode (from D2D mode) and from RFEH mode (to D2D mode), i.e., P_{RFI} and P_{RFO}, respectively. It is assumed that the plot of P_{RFI} (or P_{RFO}) vs. W_B is similar to step function where $\{\epsilon_1, \epsilon_2\}$ denotes the W_B at which there is step change in P_{RFI} and P_{D2D}, respectively where $\epsilon_1 < \epsilon_2$ and $\{\eta_1, \eta_2\}$ decides the slope of these step functions, respectively. Ideally, the total reward (or regret) of the policy should be independent of $\{\epsilon_1, \epsilon_2\}$ values (i.e., mode switching frequency), since many factors such as location of SU, priority of data transmission, available battery charge, type of service, future services and location etc. might influence the mode switching decision in practice. Intuitively, the proposed D2D-RFEH policy satisfies this requirement since the mode switching events does not affect performance of the policy. This will be verified using the simulation results presented in the Section 4.

4. PERFORMANCE ANALYSIS

In this section, we evaluate and compare the total reward and SSC of the proposed policy with the ρ^{rand} policy in [6] and its extensions. The mode switching scheme is same as Section 3.3 for all the policies. At the start, the batteries of all SUs are 50% charged. Consider N=16 (i.e., B_{cmin}=1/16), $M \in \{3,4,..12\}$ and two P_{vac} distributions (case 1[1] and case 2[2]). Each numerical result reported hereafter is the average of the values obtained over 50 independent experiments.

For M=5 and B_v={1/16, 3/16, 2/16, 2/16, 1/16}, the plots of average reward, S_t, in percentage vs. time slots, t, for case 1[1] and case 2[2] are shown in Figures 1a and 1b, respectively where $S_t \ \% = [100 \cdot (S_t^* - S_t)/S_t^*]$ and S_t^* is the average reward of the genie-aided policy. The plots in Figures 1c and 1d for case 1[1] and case 2[2], respectively, correspond to the conventional D2D mode with no need of RFEH. It can be observed that the proposed policy offers 5–25% higher reward (and hence, higher spectrum efficiency) than other policies.

Next, 50 different combinations of $M \in \{3,4,..,12\}$, $B_v \preceq (1/2)$ and $P_{vac} \in$ {case 1[1] and case 2[2]} for seven distinct scenarios are considered. Remaining parameters are same as those in case of Fig. 1. In Fig. 2 (a), the comparison is made between different policies based on the average reward, $S(t)$ in % with respect to $S(t)$ of the proposed policy. In Fig. 2 (b), the average SSC in % of different policies is compared with respect to that of UCBB+ρ^{rand} policy where UCBB corresponds to block UCB algorithm having lower SSC than conventional UCB algorithm. In summary, the total reward and SSC of the Bayesian algorithms based policies are

[1] P_{vac} :- [.05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .90 .95]
[2] P_{vac} :- [.37 .04 .13 .01 .03 .17 .05 .50 .60 .06 .02 .35 .25 .99 .20 .10]

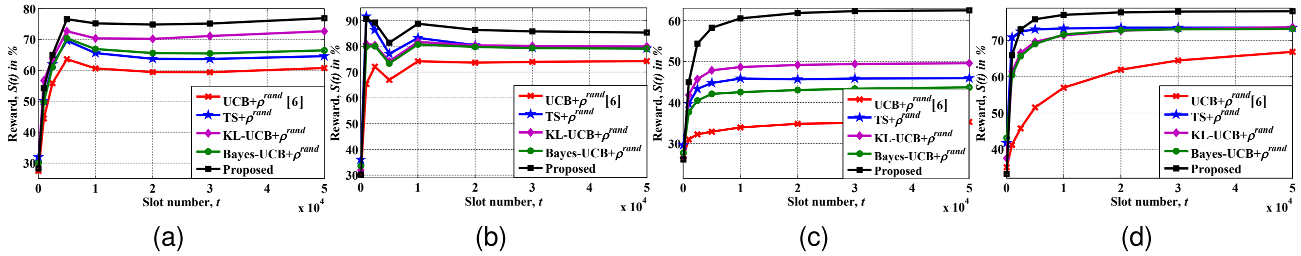

Fig. 1. Average reward, $S(t)$ in % vs. time slot, t, for (a) Case 1, (b) Case 2, in the D2D-RFEH scenario, (c) Case 1, (d) Case 2, in the conventional DSLA scenario with no need of RFEH where $\{\epsilon_1, \epsilon_2\} = \{0.3, 0.4\}$, $\{\eta_1, \eta_2\} = \{8, 70\}$, ξ=0.1 and θ=4.

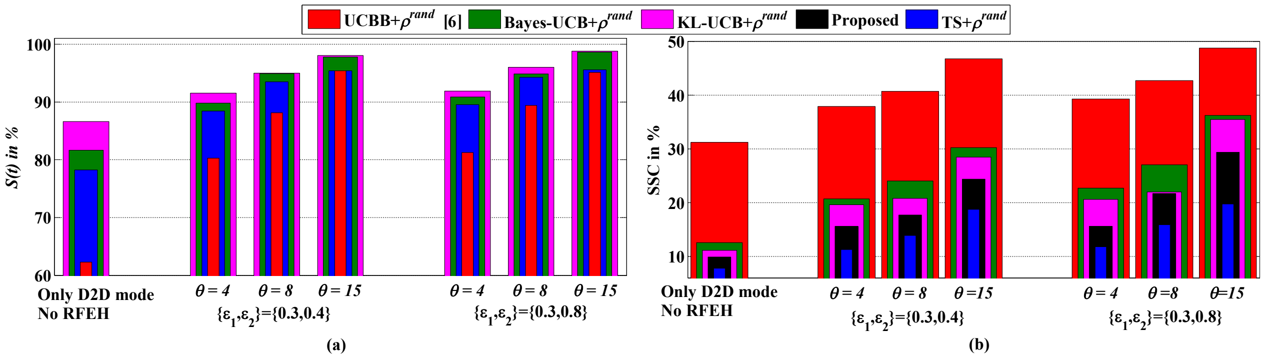

Fig. 2. (a) The plot of average $S(t)$ in % with respect to $S(t)$ of the proposed scheme, (b) The plot of average SSC in % with respect to $S(t)$ of the UCB+ρ^{rand} policy in [6].

significantly better than frequentist approach algorithm based policies. Among Bayesian policies, proposed policy not only offers superior performance but also have lower complexity.

Since there is no constraint of collisions among SUs in RFEH mode, average time spend by SUs in RFEH is nearly same for fixed bandwidth RFEH (i.e. $B_o(k) = B_{cmin}$) case. As a result, the gain in $S(t)$ decreases as θ increases. Nonetheless, it is expected that efficient RFEH circuits with $\theta < 5$ will be available in near future for which the proposed policy offers more than 10% gain in $S(t)$ over other policies.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, a new dynamic spectrum learning and tunable bandwidth access based device-to-device communications and radio frequency energy harvesting (D2D-RFEH) policy for decentralized cognitive radio networks (CRNs) is proposed. Simulations results and complexity analysis showed that the proposed policy offers superior performance over the frequentist approach based policies making it suitable for resource–constrained battery operated secondary users. Future work considers tunable bandwidth access in RFEH mode for faster RFEH thereby further improvement in reward.

REFERENCES

[1] A. Asadi, Q. Wang, and V. Mancuso, “A Survey on Device-to-Device Communication in Cellular Networks,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801-1819, Nov. 2014.

[2] A. H. Sakr and E. Hossain, “Cognitive and Energy Harvesting-based D2D Communications in Cellular Networks: Stochastic Geometry Modeling and Analysis,” *IEEE Trans. on Comm.*, April 2015.

[3] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, “Wireless Networks with RF Energy Harvesting: A Contemporary Survey,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 757-789, May 2015.

[4] D. T. Hoang, D. Niyato, P. Wang, and D. I. Kim, “Opportunistic Channel Access and RF Energy Harvesting in Cognitive Radio Networks,” *IEEE Trans. on Selected Areas in Comm.*, vol. 32, no. 11, pp. 1-14, Nov. 2014.

[5] S. Park, J. Heo, B. Kim, W. Chung, H. Wang, and D. Hong, “Optimal Mode Selection for Cognitive Radio Sensor Networks with RF Energy Harvesting,” in *Proc. IEEE* 23^{rd} *International Symposium on PIMRC*, Sydney, Australia, Sept. 2012.

[6] A. Anandkumar, N. Michael, A. Tang, and A. Swami, “Distributed Algorithms for Learning and Cognitive Medium Access With Logarithmic Regret,” *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, April 2011.

[7] K. Liu and Q. Zhao, “Distributed Learning in Multi-Armed Bandit with Multiple Players,” *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[8] S. J. Darak, H. Zhang, J. Palicot, and C. Moy, “Efficient Decentralized Dynamic Spectrum Learning and Access Scheme for Multi-standard Multi-user Cognitive Radio Networks,” in 11^{th} *IEEE Int. Symp. on Wireless Comm. Systems*, pp. 271-275, Barcelona, Spain, Aug. 2014.

[9] S. J. Darak, C. Moy, S. Dhabu, H. Zhang, J. Palicot and A. P. Vinod, “Decentralized Spectrum Learning and Access for Heterogeneous Cognitive Radio Networks,” *Digital Signal Processing (Elsevier)* , vol. 37, pp. 13–23, Feb. 2015.

[10] E. Kaufmann, O. Cappé, and A. Garivier, “On the Efficiency of Bayesian Bandit Algorithms from a Frequentist Point of View,” *Neural Information Processing Systems (NIPS)*, Dec. 2011.

[11] D. Russo and B. V. Roy, “An Information-Theoretic Analysis of Thompson Sampling, ” *Computing Research Repository*, Mar. 2014.

[12] A. Garivier and O. Cappé, “The KL-ucb Algorithm for Bounded Stochastic Bandits and Beyond, ” *Conference On Learning Theory (COLT)*, pp. 359-376, Budapest, Hungary, July 2011.