Question(s):	-	**Meeting:**	FG-ML5G meeting and workshop, 29 January – 2 February 2018
Study Group:	Focus Group on on Machine Learning for Future Networks including 5G	**Working Party:**	Department of Information and Communication Technologies, Universitat Pompeu Fabra (UPF)
Source:	-		
Title:	Implications of Decentralized Learning in Dense WLANs		
Purpose:	[Choose a purpose from the dropdown list]		

Contact:	Francesc Wilhelmi Roca Universitat Pompeu Fabra Spain	Tel: +34935422906 Fax: - E-mail: francisco.wilhelmi@upf.edu

Keywords: Next-generation wireless networks; Spatial Reuse; Decentralized learning; Multi-Armed Bandits; Thompson sampling

Abstract: Understanding the consequences of applying Reinforcement Learning (RL) in dense and uncoordinated environments (e.g., Wi-Fi) is critical to optimize the performance of next-generation wireless networks. In this document we present a decentralized approach in which Wireless Networks (WNs) attempt to learn the best possible configuration in an adversarial environment according to their own performance. In particular, we provide a Multi-Armed Bandits (MABs) based model in which devices are allowed to tune their frequency channel, transmit power and Carrier Sense Threshold (CST). Our results show that, despite using only local information, a collaborative behavior can be obtained among independent devices that share the same resources. Furthermore, we study the effects of applying such method under different equilibrium situations with respect to the adversarial setting. Finally, some insights are provided regarding the consequences of applying learning in presence of legacy nodes.

1. Introduction

Wireless communications are rapidly evolving to satisfy the increasingly tighter requirements coming from the explosive growth of wireless devices and the Quality of Service (QoS) needs of new applications. Next-generation Wireless Networks (WNs) are foreseen to cover short-range communications in high-density scenarios, which most of the cases are uncoordinated deployments (e.g., residential buildings). Because of the limited performance of the existing protocols when operating in dense scenarios, which suffer a high variability in terms of channel effect, users mobility and traffic patterns, Reinforcement Learning (RL) is gaining attention in the wireless communications community. Notwithstanding, the application of learning techniques in wireless networks entails a set of trade-offs that must be carefully considered, specially in fully decentralized scenarios. On the one hand, a limited-information regime may constrain the performance of potential learning mechanisms because an overall vision of the network is not available. On the other hand, the fact of dynamically applying a set of actions may have severe consequences on legacy nodes that overlap with the learning agents.

In this document we present a decentralized learning approach for dynamically tuning the transmit power, the Carrier Sense Threshold (CST), and the frequency channel in IEEE 802.11 Wireless Local Area Networks (WLANs), which is an extension of the work in [1]. While the application of Transmit Power Control (TPC) and dynamic CST adjustment are intended to increase Spatial Reuse (SR), Dynamic Channel Allocation (DCA) allows to relax the problem by minimizing the number of overlapping WLANs.

In overview mode, the main contributions done here are described next:

- We model the SR problem through Multi-Armed Bandits (MABs), which frames the exploration-exploitation dilemma for unknown scenarios, and allows to remove the complexity added by a states-based model.
- We show that decentralized learning in adversarial wireless networks leads to collaborative results, even if no information is shared.
- We showcase the implications of applying selfish learning in an adversarial setting with different equilibrium situations, which depend on the spatial distribution of the WLAN.
- We give insight on the effects of learning selfishly in presence of legacy nodes.

The remaining of this document is structured as follows: Section 2 introduces to the previous related work regarding RL application in wireless networks. The system model is presented in Section 3, where the proposed decentralized approach is presented. Then, Section 4 shows the main results of this proposal. Finally, some remarks are given in Section 5.

2. Related Work

Machine Learning (ML), and more precisely Reinforcement Learning, has received increasing interest from the wireless communications research community over the last years. The main reason lies in the increased complexity of problems related to next-generation wireless systems, which are characterized by being dense and varying. Henceforth, RL helps at approximating solutions to complex problems that cannot be solved within an acceptable timescale. Furthermore, a high variability of the environment (e.g., nodes location, traffic generation) compromises the validity of the previously collected information about the system. Thus, online learning stands as an appropriate framework to allow performance maximization in wireless systems.

To the best of our knowledge, one of the first works to apply RL in a resource allocation problem in wireless networks is [2], in which the authors propose a real-time Q-learning mechanism to dynamically select the channel in mobile networks. To model the problem, it is assumed that the wireless environment is a discrete-time event system, so that learning is sequentially performed in cells in which an event occurred. According to that premise, a state s_t at time t is considered to include the cell index i in which an event occurred (e.g., a user arrival), and the set of available channels $A(i)$ from i's perspective, which depends on the sensed interference at time t. Furthermore, applying an action consists in selecting a channel from $A(i)$. Finally, the reward experienced by the overall system is computed as a function of the channel utilization. The application of Q-learning is however considered to be done in a central or a distributed system, but the overhead costs are not considered. Q-learning has been also applied for channel selection in wireless networks in [3-7]. While [3-5] focus on channel allocation approaches, [6, 7] also introduce transmit power adjustment.

Despite of the recent popularity of RL, and more specifically Q-learning, for its application in wireless networks, there are still many problems that are not properly modeled through such methods. The fact of modeling states and actions is not always feasible and practical, especially in adversarial environments. For that, we focus our attention on the Multi-Armed Bandits (MABs) problem, in which states are not considered in general. Instead of using state-action pairs to devise an optimal policy, in MABs the main goal is to learn the hidden reward distribution of actions in order to exploit them as efficiently as possible. MABs are used to model the channel selection and power control problem in Device-to-Device (D2D) networks in [8, 9]. In such context, the behavior of D2D users (each one composed by a transmitter and a receiver) is studied under the adversarial setting. While [8] aims to study the performance of different learning strategies, [9] includes a calibrated predictor (referred as *forecaster*) to infer the behavior of the other devices in order to counter act their actions. In both cases, results show an optimal resource allocation in terms of channel sharing without adding any kind of communication. Both works rely in the existence of a unique pure-strategy Nash Equilibrium, which favors convergence.

However, there are many problems in which the application of MABs is not straightforward, especially in dense and unknown environments in which many devices share the same channel resources and their interactions are hard to be modeled. Such adversarial setting increases the system complexity and prevents to provide an accurate states model, which may harm the algorithm's performance. This issue is studied in [10], which defends the importance of exploiting dependencies between actions in order to efficiently solve combinatorial optimization problems. In this context, the authors in [11] approach more challenging scenarios through MABs. The main presented problem refers to dynamic rate and channel selection at stationary cognitive radio systems, but non-stationary environments are also considered for further extensions. In particular, structured MABs are used to model the problem, thus taking advantage of the correlations between different rates with regard to packet losses. Such problem modeling allows to provide a clever exploration strategy that finds the best channel-rate pair quickly and efficiently.

To the matter of this paper, we focus attention on the work in [1, 12], which emphasize on the consequences of decentralized learning in adversarial environments, which are typically found in IEEE 802.11 WLANs (e.g., residential buildings, shopping malls). For that purpose, [1] presents a channel allocation and power control approach through an stateless variation of Q-learning, so that individual WNs attempt to learn the best configuration according to their own performance, i.e., using only local information. However, the application of RL in a fully decentralized environment is shown to generate an adversarial setting that severely affects to the performance variability. Such phenomena is more evident if limited resources are shared. On the other hand, the work in [12] proposes a further analysis of the problem by approaching it through MABs, which shows better results in terms of variability. Among the action-selection strategies presented, we highlight Thompson sampling, which allows fast convergence at solving the problem even in the presence of adversarial nodes. An important remark is that collaborative learning is shown to be feasible even if selfish strategies are applied in fully decentralized environments. Decentralized learning has been also pointed out for its practical application in channel access problems in wireless networks [13, 14].

3. System Model

Next, we describe the SR reuse problem modeling through MABs, and more specifically when applying Thompson sampling. Additionally, we describe the main assumptions done to characterize a wireless scenario, which are key to understand the interactions that occur between devices.

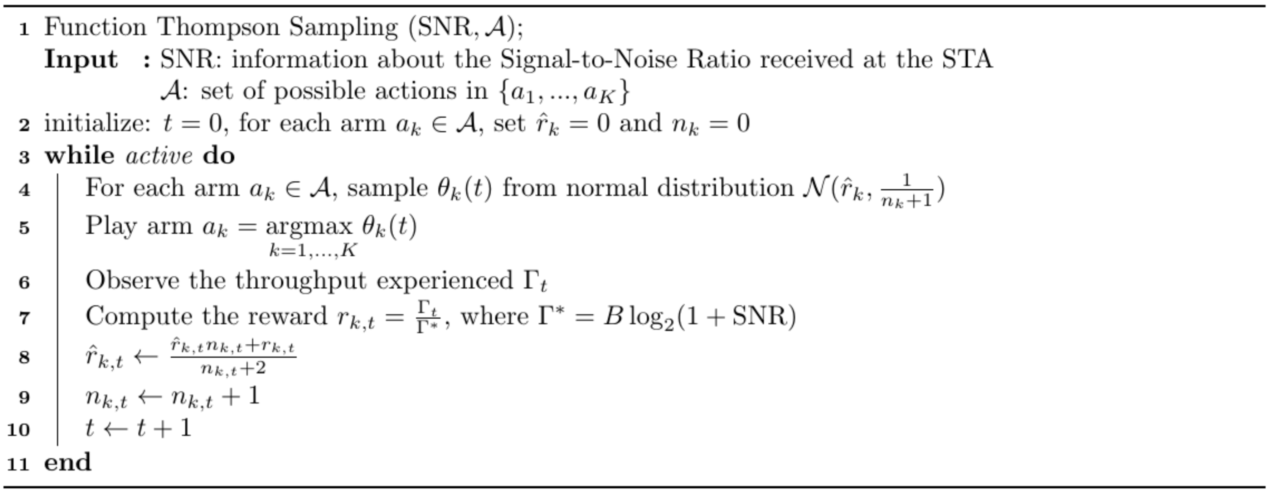
3.1 Multi-Armed Bandits to Enhance Spatial Reuse in WLANs

The SR problem in wireless networks is modeled through MABs in a decentralized manner. Such adversarial problem has been previously modeled through MABs in [12], so that learning devices make actions at specific time intervals, i.e., iterations. We refer to the period between two iterations as an unspecified time slot in which a WLAN is able to obtain an accurate measure of its actual performance (e.g., 10 minutes). In our model, assuming that all the WLANs are learning agents, actions are made simultaneously at the beginning of each iteration, so that their impact can be evaluated at the end of the latter.

Regarding the action-selection strategy, we employ Thompson sampling, which has been shown to grant excellent performance in front of other well-known policies such as UCB or EXP3 when applied in wireless networks [12]. Thompson sampling [15] is a Bayesian algorithm that constructs a probabilistic model of the rewards and assumes a prior distribution of the parameters of said model. Given the data collected during the learning procedure, Thompson sampling keeps track of the posterior distribution of the rewards, and pulls arms randomly in a way that the drawing probability of each arm matches the probability of the particular arm being optimal. In practice, this is implemented by sampling the parameter corresponding to each arm from the posterior distribution, and pulling the arm yielding the maximal expected reward under the sampled parameter value. For the sake of practicality, we assume that rewards follow a normal distribution, as suggested in [16]. By standard calculations, it can be verified that the posterior distribution of the rewards under this model is Gaussian with mean $\hat{r}_k(t) = \frac{x \sum_{w=1:k}^{t-1} r_k(t)}{n_k(t)+1}$ and variance $\sigma_k^2(t) = \frac{1}{n_k+1}$, where n_k is the number of times that arm k was drawn until the beginning of round t. Thus, implementing Thompson sampling in this model amounts to sampling a parameter θ_k from the Gaussian distribution $\mathrm{N}(\hat{r}_{k,t}, \sigma_k^2(t))$ and choosing the action with the maximal parameter. Roughly, each WLAN applies the Thompson sampling strategy as follows:

- Initially, the estimate of each action $k \in \{1, \ldots, K\}$ is set to 0.
- In each iteration, it pulls an arm randomly according to the generated probabilistic model, so that the action with the highest drawn estimate is chosen, $\hat{r}_{k,t}$.
- After choosing action a_k, it observes the reward generated by the environment, $r_{k,t}$, which also depends on the actions made by the overlapping WLANs.
- At the end of the iteration, the estimated reward of each action is updated according to the actual experienced reward and the number of times an action has been played so far: $\hat{r}_{k,t+1} = \frac{\hat{r}_{k,t}\, n_{k,t} + r_{k,t}}{n_{k,t}+2}, \forall\, k \in \{1, \ldots, K\}$.

In the decentralized setting, the reward is computed as the normalized throughput, i.e., the throughput experienced by the WLAN, divided by the optimal throughput that it would achieve in isolation. Our implementation of Thompson sampling to the WLAN problem is detailed in Algorithm *1*.



```
Function Thompson Sampling (SNR, $\mathcal{A}$);
Input : SNR: information about the Signal-to-Noise Ratio received at the STA
        $\mathcal{A}$: set of possible actions in $\{a_1, ..., a_K\}$
initialize: $t = 0$, for each arm $a_k \in \mathcal{A}$, set $\hat{r}_k = 0$ and $n_k = 0$
while active do
    For each arm $a_k \in \mathcal{A}$, sample $\theta_k(t)$ from normal distribution $\mathcal{N}(\hat{r}_k, \frac{1}{n_k+1})$
    Play arm $a_k = \underset{k=1,\ldots,K}{\operatorname{argmax}}\ \theta_k(t)$
    Observe the throughput experienced $\Gamma_t$
    Compute the reward $r_{k,t} = \frac{\Gamma_t}{\Gamma^*}$, where $\Gamma^* = B \log_2(1 + \mathrm{SNR})$
    $\hat{r}_{k,t} \leftarrow \frac{\hat{r}_{k,t} n_{k,t} + r_{k,t}}{n_{k,t}+2}$
    $n_{k,t} \leftarrow n_{k,t} + 1$
    $t \leftarrow t + 1$
end
```

Algorithm 1: Implementation of Multi-Armed Bandits (Thompson sampling) in a WLAN

3.2 Channel and Throughput Calculation Models

Path-loss and shadowing effects are modeled through the IEEE 802.11ax residential scenario [17], which is representative for next-generation dense and chaotic deployments. The path-loss PL_d in such a scenario is given by:

$$PL_d = 40.05 + 20\log_{10}\left(\frac{f_c}{2.4}\right) + 20\log_{10}(\min(d,5)) + I_{\{d>5\}} 35\log_{10}\left(\frac{d}{5}\right) + 18.3\ F^{\frac{F+2}{F+1}-0.46} + 5W$$

where f_c is the frequency in GHz, d is the distance between the transmitter and the receiver in meters, and F and W are the average number of floors and walls traversed per meter, respectively. Regarding adjacent channel interference, we consider that consecutive channels are non-overlapping.

In order to focus on the inter-WLAN interactions, and for the sake of simplicity, we consider that each WLAN is composed by a single AP and a STA, so that only the AP acts as a transmitter. Accordingly, the performance experienced by each WLAN depends on the power received at the STA from its AP and the sensed interference. In particular, to calculate the throughput experienced by each WLAN, we use the CTMN model [18], which allows to capture the Distribution Coordination Function (DCF) applied by IEEE 802.11 WLANs. Using the CTMN model allows us to obtain transparent results to properly understand the behavior of WLANs, since such analytical model has been widely used by the research community. In a CTMN, the throughput experienced by each node can be computed by measuring the time a node is allowed to transmit in presence of overlapping devices.

4. Performance Evaluation

In this Section we present the main results obtained from a set of simulations[1], which aim to shed some light on the impact of applying RL in different equilibrium situations. Simulations make use of the SF-CTMN framework presented in [19].

4.1 Scenario and System Parameters

In order to draw very concrete conclusions regarding the implications of learning in wireless networks, we consider a symmetric scenario containing a grid of 4 WLANs (shown in Figure *1*), each one formed by an AP and a STA, so that data transmissions are carried out in the downlink.

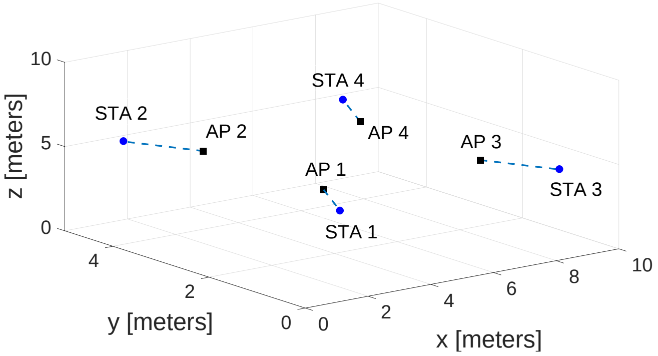


Figure 1: Simulation scenario

Such scenario is very illustrative to our purposes due to the equilibriums that can be achieved by granting slight modifications regarding the spatial distribution and the possible actions. The situations of interest are described next:

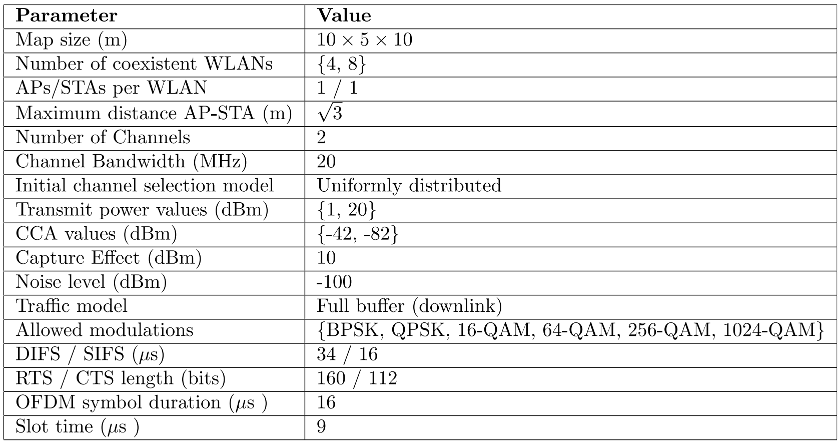
- **Weak Equilibrium (WE):** there is a unique solution that maximizes both the aggregate throughput and the proportional fairness (all the WLANs can access to the channel simultaneously), which entails that all the WLANs use the lowest transmit power and the highest CCA. While the former allows reducing the generated interference, the latter is useful to reduce the listening area. Despite such a solution leads to an equilibrium (any deviation of the profile results in a lower throughput and may affect the other players), it is very likely to fall into a suboptimal Nash Equilibrium (NE) in which all the WLANs use the maximum transmit power (or alternatively, the minimum CCA) and listen each other. We refer to this situation as a weak equilibrium. In particular, reaching the optimal NE solution requires that all the WLANs play the highest-rewarding action at a time, since any deviation of the strategy profile leads to a suboptimal configuration (similar to the prisoner's dilemma). For instance, if WLAN A decreases its transmit power in order to reach the optimal global solution, but WLAN B still uses the maximum transmit power, an imbalance is created due to the asymmetry between A and B. Henceforth, B occupies

[1] The source code and other related documentation regarding this project are available at the following online repository: https://github.com/fwilhelmi/implications_of_decentralized_learning_in_dense_wlans

all the channel without having to decrease its transmit power, while A experiences a null performance. Since using the minimum transmit power is risky and may lead to obtain a payoff of 0, the dominant NE strategy entails using the maximum transmit power.

- **Strong Equilibrium (SE):** as for WE, there is an optimal solution that maximizes both aggregate throughput and proportional fairness. In contrast, reaching such a solution becomes easier because it is built by a dominant pure-strategy NE. To do so, we provide a higher maximum CCA value, namely CCA$_{\text{max}}$. Thus, the optimal action, a*, that maximizes both aggregate throughput and proportional fairness is strong enough to grant maximum individual performance despite the configuration of the other nodes.
- **Overlapping Strong Equilibrium (OSE):** there is not an optimal solution that maximizes both aggregate throughput and proportional fairness. While maximizing the former requires that one WLAN per channel obtains full access to the medium, the later entails a full overlapping situation in which each frequency channel is fairly shared over the time. To force that situation we place nodes a bit closer between them. Henceforth, unlike for WE, the previously seen optimal solution drives to null performance due to collisions by hidden node (capture effect is not considered). With that, we illustrate the adversarial setting in which WLANs must settle for a portion of the available bandwidth.

The simulation parameters, which include IEEE 802.11ax PHY and MAC specifications [20], are shown in *Table 1*.



Parameter	**Value**
Map size (m)	10 × 5 × 10
Number of coexistent WLANs	{4, 8}
APs/STAs per WLAN	1 / 1
Maximum distance AP-STA (m)	$\sqrt{3}$
Number of Channels	2
Channel Bandwidth (MHz)	20
Initial channel selection model	Uniformly distributed
Transmit power values (dBm)	{1, 20}
CCA values (dBm)	{-42, -82}
Capture Effect (dBm)	10
Noise level (dBm)	-100
Traffic model	Full buffer (downlink)
Allowed modulations	{BPSK, QPSK, 16-QAM, 64-QAM, 256-QAM, 1024-QAM}
DIFS / SIFS (μs)	34 / 16
RTS / CTS length (bits)	160 / 112
OFDM symbol duration (μs)	16
Slot time (μs)	9

Table 1: Simulation parameters

4.2 Optimal Solution

Before presenting our main results it is important to show, for each equilibrium scenario, the optimal solutions that *(a)* maximize the aggregate throughput and *(b)* correspond to proportional fairness. The proportional fairness solutions satisfy $\max_{a_k \in A} \sum_{i \in WLAN} \log_{10} \Gamma_{i,a_k}$.

Let actions range from a_1 to a_8, and correspond to the following combinations of {channel number, CCA (dBm), transmit power (dBm)}: a_1 = {1, -42, 1}, a_2 = {2, -42, 1}, a_3 = {1, -82, 1}, a_4 = {2, -82, 1}, a_5 = {1, -42, 20}, a_6 = {2, -42, 20}, a_7 = {1, -82, 20} and a_8 = {2, -82, 20}. Then, the configuration that grants the maximum aggregate throughput in WE (531.27 Mbps) is any combination of {a_2, a_1, a_1, a_2} in which diagonal nodes use the same configuration. Such a configuration also provides the proportional fair solution, since there exists a Nash Equilibrium when WLANs follow a pure strategy. Additionally, in SE, the optimal configuration is the one that uses the maximum transmit power and CCA threshold, i.e., 20 dBm and -32 dBm, respectively. For instance, { a_7 , a_2 , a_6 , a_5 }. In contrast, for OSE, the optimal aggregate throughput and proportional fair solutions do not match. For the former, the maximum aggregate throughput (267.24 Mbps) is granted by any consecutive combination of {a_5, a_6, a_4, a_3}, in which two of the WLANs (using different channels) enjoy much more throughput than the other ones. Furthermore, the proportional fair result (266.84 Mbps) is provided by any combination of

$\{a_6, a_5, a_5, a_6\}$ and $\{a_4, a_3, a_3, a_4\}$. Moreover, in this scenario we obtain a null performance (0 Mbps) if applying any combination of $\{a_8, a_7, a_7, a_8\}$, since all the WLANs experience a very high collisions rate because the CCA condition is met, but the CE is not.

Note, as well, that in any of the presented equilibrium situations, multiple combinations of a given configuration lead to the optimal solution due to the symmetry of the scenario.

4.3 Decentralized Learning Results

Here we show the decentralized case in which WLANs use the individual throughput as a reward, which was the main case of study in [1]. We run simulations of 1,000 iterations.

Figure 3 shows the probability of choosing an action for each WLAN, and for each of the equilibrium scenarios (SE, WE and OSE). For the WE situation, we observe that a larger range of actions is chosen than in SE, in which the optimal action is rapidly found. Regarding OSE, similar behavior than for WE is shown.

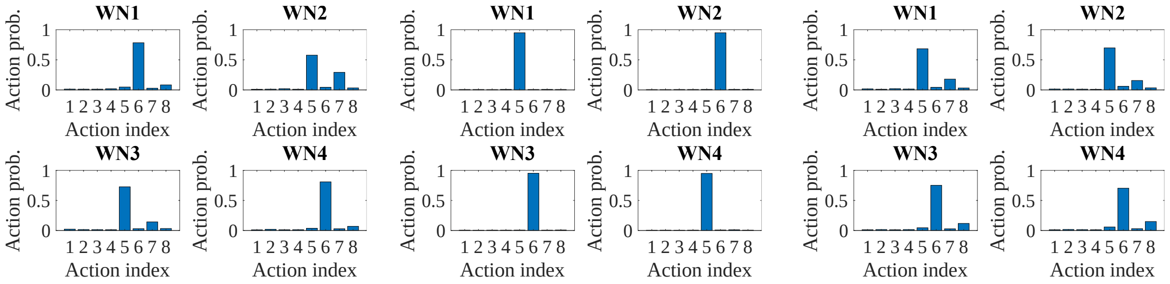


Figure 2: Actions probability per WLAN in each scenario (left: WE, center: SE, right: OSE)

We can further analyze the performance of Thompson sampling in each scenario in Figure 4, Figure 5 and Figure 6, which show the mean throughput experienced per WLAN after applying Thompson sampling for 1.000 iterations, and the aggregate and individual throughput evolution, respectively. From Figure 6 we can observe that a fast convergence is achieved to the optimal fair solution at SE and OSE scenarios, which at the same time maximizes the aggregate throughput (thus providing a 100% system efficiency). Regarding WE, convergence to the optimal solution is not achieved, but a fair configuration is granted.

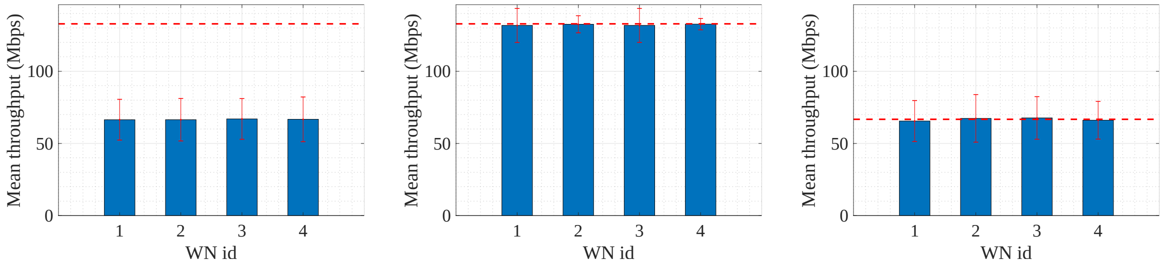


Figure 3: Average throughput per WLAN in each scenario (left: WE, center: SE, right: OSE)

Regarding the temporal throughput variability, we see that it is especially intensive in the OSE case, in which WLANs must share the channel and renounce to their maximum individual performance. In such a competitive case, although system efficiency is not maximum, the proportional fair solution is achieved through decentralized learning.

Similarly, in the WE case, WLANs experience a high variability because their obtained solution does not match with the optimal one. We identify the cause of this variability to be the larger range of actions that grant high individual performance, but that harm the overall fairness (e.g., configurations that use a high transmit power and a high CCA). Finally, regarding the SE situation, we notice a low variability and a fast convergence towards the optimal solution.

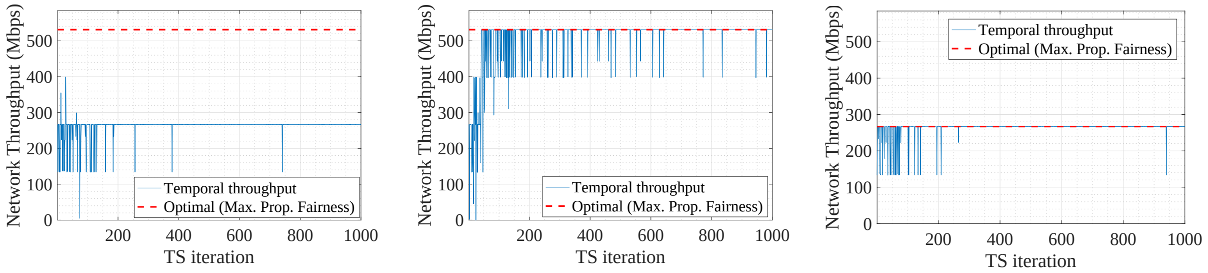


Figure 4: Temporal aggregate throughput in each scenario (left: WE, center: SE, right: OSE)

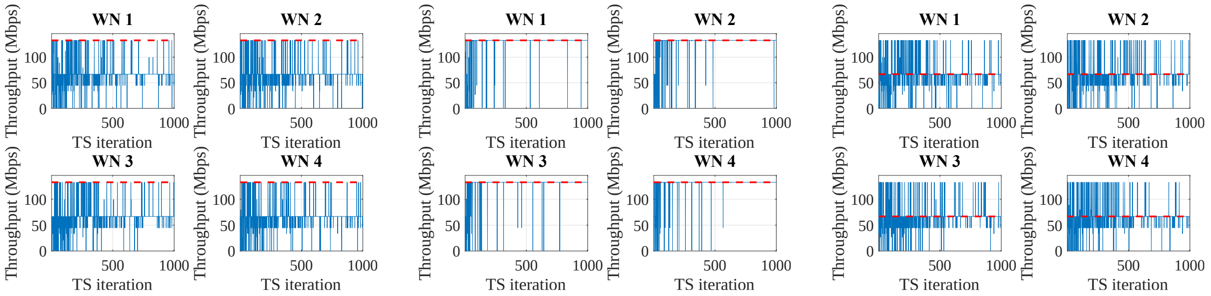


Figure 5: Temporal throughput per WLAN in each scenario (left: WE, center: SE, right: OSE)

4.4 Implications of decentralized learning in legacy devices

Finally, in order to further analyze the implications of applying decentralized learning, we propose a random scenario containing 8 WLANs (each one formed by an AP-STA pair), in which Thompson sampling is applied selfishly in presence of legacy devices. In particular, the random scenario is maintained for all the following simulations, but the percentage of WLANs that are legacy varies. For each legacy percentage (0, 25, 50 and 75%), WLANs are randomly determined to include learning agents or not.

In Figure *6* we show the average throughput obtained by each WLAN class (i.e., legacy and learning) for each Thompson sampling iteration. Results are shown for each percentage of legacy devices in the network, so that we are able to study the impact of learning according to the magnitude of legacy nodes. As it is shown, the performance of legacy devices is severely affected by the actions done by learning devices, which is clearly observed in the 25% case, where legacy's average performance falls to 0. The fact that learning devices employ more aggressive configurations entails a lack of fairness with respect to legacy WLANs. Another important consideration lies in the throughput variability, which becomes lower as the number of legacy devices increases, i.e., for more stable environments.

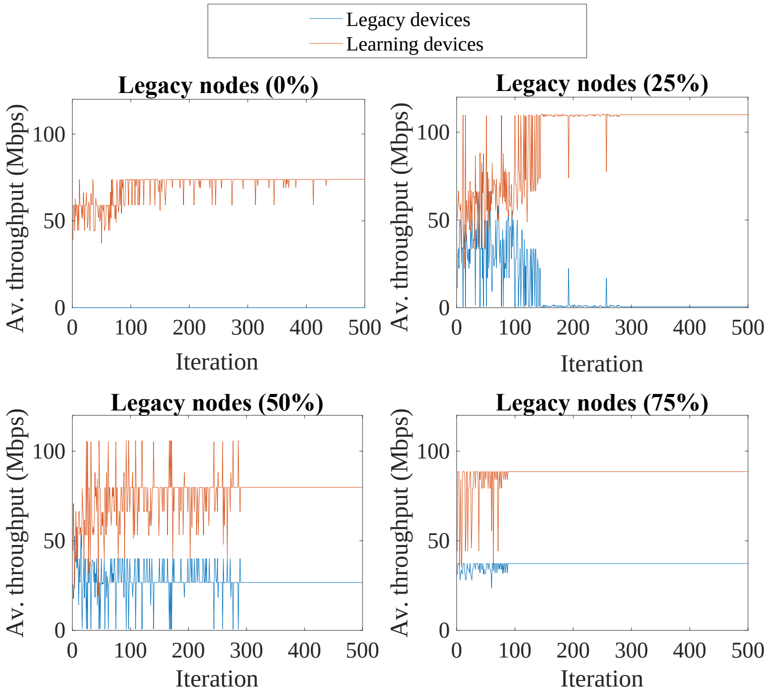


Figure 6: Average throughput experienced by legacy (blue) and non-legacy nodes (red).

5. Conclusions

In this document we presented a decentralized learning mechanism to improve the performance of IEEE 802.11 WLANs through SR maximization, and to study the main derived implications of learning in adversarial environments. For that, we presented three different equilibrium situations, and show that learning WLANs show a collaborative behavior in which fairness is maximized. However, in non-strong equilibrium situations the optimal solution cannot be achieved, which causes variability issues in terms of experienced throughput, which may severely affect to upper the communication layers (e.g., congestion window procedure in TCP). Furthermore, we studied the effect of applying decentralized learning in presence of legacy nodes (i.e., static devices), and showed that the performance of the latter may be severely harmed.

Acknowledgement

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), and by a Gift from CISCO University Research Program (CG#890107) & Silicon Valley Community Foundation.

Bibliography

[1] Wilhelmi, F., Bellalta, B., Cano, C., & Jonsson, A. (2017). Implications of Decentralized Q-learning Resource Allocation in Wireless Networks. arXiv preprint arXiv:1705.10508.

[2] Nie, J., & Haykin, S. (1999). A Q-learning-based dynamic channel assignment technique for mobile communication systems. IEEE Transactions on Vehicular Technology, 48(5), 1676-1687.

[3] Li, H. (2009, October). Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case. In Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on (pp. 1893-1898). IEEE.

[4] Sallent, O., Pérez-Romero, J., Ferrús, R., & Agustí, R. (2015, June). Learning-based coexistence for LTE operation in unlicensed bands. In Communication Workshop (ICCW), 2015 IEEE International Conference on (pp. 2307-2313). IEEE.

[5] Rupasinghe, N., & Güvenç, İ. (2015, March). Reinforcement learning for licensed-assisted access of LTE in the unlicensed spectrum. In Wireless Communications and Networking Conference (WCNC), 2015 IEEE (pp. 1279-1284). IEEE.

[6] Bennis, M., & Niyato, D. (2010, December). A Q-learning based approach to interference avoidance in self-organized femtocell networks. In GLOBECOM Workshops (GC Wkshps), 2010 IEEE (pp. 706-710). IEEE.

[7] Bennis, M., Guruacharya, S., & Niyato, D. (2011, December). Distributed learning strategies for interference mitigation in femtocell networks. In Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE (pp. 1-5). IEEE.

[8] Maghsudi, S., & Stańczak, S. (2015). Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework. IEEE Transactions on Vehicular Technology, 64(10), 4565-4578.

[9] Maghsudi, S., & Stańczak, S. (2015). Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. IEEE Transactions on Wireless Communications, 14(3), 1309-1322.

[10] Gai, Y., Krishnamachari, B., & Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. IEEE/ACM Transactions on Networking (TON), 20(5), 1466-1478.

[11] Combes, R., & Proutiere, A. (2015). Dynamic rate and channel selection in cognitive radio systems. IEEE Journal on Selected Areas in Communications, 33(5), 910-921.

[12] Wilhelmi, F., Cano, C., Neu, G., Bellalta, B., Jonsson, A., & Barrachina-Muñoz, S. (2017). Collaborative Spatial Reuse in Wireless Networks via Selfish Multi-Armed Bandits. arXiv preprint arXiv:1710.11403.

[13] Liu, K., & Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. IEEE Transactions on Signal Processing, 58(11), 5667-5681.

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

[15] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25, 285-294.

[16] Agrawal, S., & Goyal, N. (2013, April). Further optimal regret bounds for thompson sampling. In Artificial Intelligence and Statistics (pp. 99-107).

[17] Merlin, S. et al. (2015). TGax simulation scenarios. IEEE.

[18] Bellalta, B., Zocca, A., Cano, C., Checco, A., Barcelo, J., & Vinel, A. (2014). Throughput analysis in CSMA/CA networks using continuous time Markov networks: a tutorial. In Wireless Networking for Moving Objects (pp. 115-133). Springer International Publishing.

[19] Barrachina-Muñoz, S., Wilhelmi, F., & Bellalta, B. (2018). Performance Analysis of Dynamic Channel Bonding in Spatially Distributed High Density WLANs. arXiv preprint arXiv:1801.00594.

[20] TGax. (2016). P802.11ax/D1.0. IEEE.