Partner selection in self-organised wireless sensor networks for opportunistic energy negotiation: A multi-armed bandit based approach

Andre P. Ortega[a,b,*], Sarvapali D. Ramchurn[a], Long Tran-Thanh[c] and Geoff V. Merrett[a]

[a]School of Electronics and Computer Science, University of Southampton, University Road, Southampton SO17 1BJ, UK
[b]Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Campus Gustavo Galindo Km 30.5 Vía Perimetral, Guayaquil, Ecuador
[c]Department of Computer Science, University of Warwick, 6 Lord Bhattacharyya Way, Coventry CV4 7EZ, UK

ARTICLE INFO

Keywords:
Wireless sensor networks
Agent-based sensor network
Energy management
Multi-armed bandit based learning
Reinforcement Learning
Automated negotiation

ABSTRACT

The proliferation of "Things" over a network creates the Internet of Things (IoT), where sensors integrate to collect data from the environment over long periods of time. The growth of IoT applications will inevitably involve co-locating multiple wireless sensor networks, each serving different applications with, possibly, different needs and constraints. Since energy is scarce in sensor nodes equipped with non-rechargeable batteries, energy harvesting technologies have been the focus of research in recent years. However, new problems arise as a result of their wide spatio-temporal variation. Such a shortcoming can be avoided if co-located networks cooperate with each other and share their available energy. Due to their unique characteristics and different owners, recently, we proposed a negotiation approach to deal with conflict of preferences. Unfortunately, negotiation can be impractical with a large number of participants, especially in an open environment. Given this, we introduce a new partner selection technique based on multi-armed bandits (MAB), that enables each node to learn the strategy that optimises its energy resources in the long term. Our results show that the proposed solution allows networks to repeatedly learn the current best energy partner in a dynamic environment. The performance of such a technique is evaluated through simulation and shows that a network can achieve an efficiency of 72% against the optimal strategy in the most challenging scenario studied in this work.

1. Introduction

In recent years, Wireless Sensor Networks (WSNs) have become an important technology for real-world environmental monitoring. The current trend to adopt the standard 6LoWPAN/IPv6 for IP-based sensor networks enables the integration of WSN applications into the Internet of Things (IoT). As a result, the likelihood of multiple WSNs being deployed in the same geographic area is expected to increase even more in the near future. Fig. 1 shows an example of a scenario with multiple sensor networks applications.

A typical WSN is composed of low-power sensing nodes with constrained power supply. Despite its potential as a perpetual energy source, energy harvesting technologies are sensitive to the intermittency inherent in some power sources (e.g. solar, wind or heat). To address this issue, we introduced a negotiation-based cooperation model for the energy harvesting wireless sensor networks (EHWSNs) [1]. Our approach allows each node to adaptively satisfy its load while it agrees to share its harvested energy at some points in time in return for energy at other points in time.

The key goal for the cooperation between EHWSNs is the efficient management of energy to enable the networks' continuous operation, also known as energy neutrality. Thus, when two or more WSNs are deployed in the same location, this work envisages a long-term cooperation between nodes. Such cooperation starts with energy agreements over a period of time in order to satisfy as much as possible the nodes' energy consumption profile.

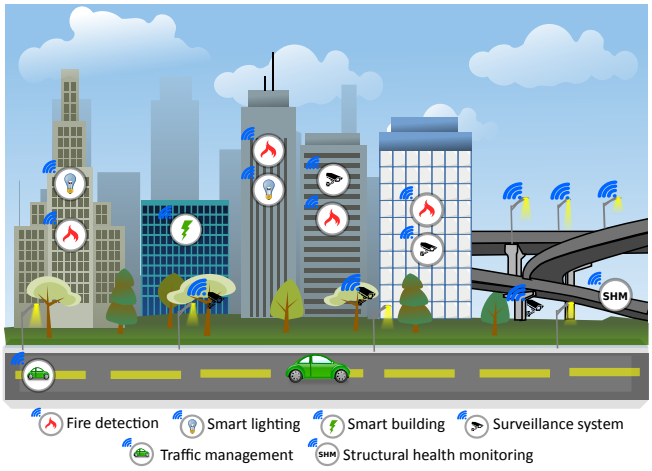

Figure 1: An IoT ecosystem with multiple co-located WSNs.

Since we are dealing with different networks that have unique characteristics and different owners, the networks can be considered independent and self-interested. Therefore, before cooperation can be decided, networks should be able to exchange offers and find a mutually-acceptable energy flow that maximises their own benefits. Accordingly, our previous work proposed a novel cooperation model based on heuristic negotiation to facilitate **O**pportunistic **E**nergy **N**egotiation between neighbouring nodes (*OEN*).

In OEN, a valid energy flow offer must include the energy values for the predetermined time of cooperation. Since

*Corresponding author
E-mail address: apoa1g15@soton.ac.uk (A.P. Ortega)

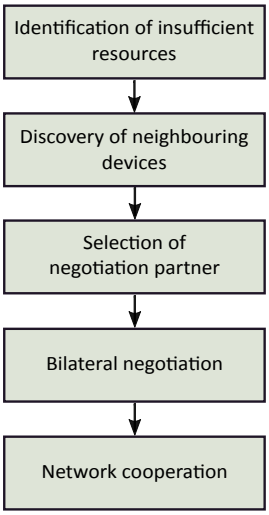

Figure 2: The 5 phases of the methodology proposed in this work to establish OEN between distinct networks.

the energy allocated on a given time slot is highly dependent on the amount received or offered in the previous slot due to the battery's dynamics, the networks bargain over interdependent issues. In the initial phase of our research, decision functions that rely on acceptable ranges of energy amounts were used to compute offers. However, the requirement of valid energy intervals increases as the number of issues in conflict increases. The approach described in this paper is an extension to the case where multiple nodes aim to participate in a negotiation and the process is an improvement of our previous methodology.

Although the cooperation seeks to optimise a system-wide goal, every single node has a limited view about the state of the entire network. This bounded knowledge is either caused by its location or constrained nature. Therefore, to optimise the network operability, the nodes must coordinate their actions with the nodes in close proximity. A multi-agent approach is a natural fit for this setting, where each sensor is controlled by an agent. The agent engages in communication with others in order to achieve system-wide goals in a distributed manner. In the same way, nodes need to adjust to topology changes, varying environmental conditions and multiple negotiation behaviours. To fully support OEN, an agent foresees an insufficient energy allocation scheme and starts the process towards cooperation with a neighbouring network. It first discovers all available agents and selects a negotiation partner from the set of co-located nodes to start a bilateral negotiation. Fig. 2 gives a general overview of our methodology.

In relatively small environments, a self-interested agent can reach its most preferred deal by negotiating with all agents that offer cooperation. The agent then chooses the most suited option to satisfy its negotiation preferences among a set of bargainers. In this situation, an agent supports robust negotiating strategies that result in efficient agreements even in dynamic environments. However, this mechanism may not be reasonable in domains with limited computation or restricted communication bandwidth. In an open dynamic domain, such as WSNs, a negotiation may lead to a communication overhead when coordinated over a large number of agents. Against this background, it is always preferable to start a negotiation which is likely to succeed and reach a better agreement. Therefore, for practical purposes, an agent should be able to anticipate the best potential negotiation partner to maximise its energy allocation.

To realise this approach, the agent seeking a partner first needs to learn the performance of all the neighbouring agents expected to cooperate. Then, the decision of the negotiation partner involves a trade-off: the negotiation with an opponent provides feedback about its effectiveness (exploration), but the collection of that feedback ignores the immediate benefit of selecting a partner already known as effective (exploitation). Within the partner selection problem, this work focuses on finding an efficient learning method to balance exploration with exploitation. Since agents have to negotiate with incomplete knowledge about their opponents and have no control of the environmental factors affecting the outcomes of the negotiation, it is very difficult to estimate the probability of the payoff structure for the partnership.

Accordingly, we suggest a light-weight learning method to address the problem of partner selection. Specifically, the partner selection can be naturally modelled as a multi-armed bandit (MAB) problem. In MABs, an agent selects from a group of known actions, the one that is most likely to yield the highest reward. The goal of the agent is to maximise the total rewards earned through a sequence of continuous observations. Our interest lies in analysing the performance of MABs algorithms when applied to WSNs. In this setting, the reward corresponds to the negotiation outcome which not only depends on the actions taken by the agent but also on the adversaries behaviour. Given this reward, an agent is able to characterise the performance of the selected opponent to solve the partner selection problem.

Unlike classical stochastic bandits, whereby the rewards are independently drawn following a fixed but unknown distribution, we study the implications of selecting an action in an adversarial setting [2]. The applicability of our solution also depends on the type of adversary. If the adversary chooses the reward ahead of the actual selection process, it is known as an oblivious adversary. Whereas, if the opponent simultaneously chooses the reward with the agent's choice, an adversary is called a non-oblivious adversary. The second category describes our case.

The contributions of this work are summarised as follows:

- A negotiation-based methodology to enable opportunistic and direct cooperation between co-located networks, where each node is able to reason if cooperation is beneficial in a decentralised manner.
- A partner selection technique based on MABs, which satisfies the simplicity and computational efficiency demanded in our domain. In particular, a comparison of four state-of-the-art algorithms is presented.

- Experimental evaluation of a generic framework for automated multi-issue negotiation applied in the context of opportunistic energy cooperation.
- An energy management method that provides a sensor node with the ability to detect an insufficient energy scheme.
- Extensive simulations to capture the energy cost of discovering nearby cooperators. The overhead of the discovery protocol is analysed in terms of the number of agents participating in the process.

The remainder of this paper is organised as follows. Section 2 gives an overview of relevant related work. Section 3 provides the system model, energy allocation algorithm and negotiation results using OEN, and defines the problem on partner selection. Section 4 demonstrates how decision-making on partner selection can be done applying MAB algorithms. Section 5 describes our experimental setting and presents the simulation results with the comparison of the MAB algorithms in every scenario. Section 6 details the establishing of OEN and its energy cost. Finally, Section 7 presents concluding remarks along with directions for future research.

2. Related work

Before presenting the related work in the problem of partner selection, we first need to introduce a review that synthesises the research on energy management, heuristic negotiation strategies, and the most relevant work for the application of reinforcement learning in the area of WSNs.

2.1. Energy management in EHWSNs

The work in energy management is studied here from the perspective of optimisation regarding energy use in rechargeable sensor networks. Previous efforts have concentrated on two different approaches of how to ensure energy-neutral sensing systems.

In this respect, [3] considers the problem of optimal power management and presents a linear programming algorithm for the adaptation of duty cycles. Similarly, work in [4] develops an energy allocation algorithm for the optimal use of energy harvested in WSNs, but the objective is to minimise the variation in allocated energy over a period of time. Both algorithms focus on the optimisation of power management when nodes are harvesting-aware but differ greatly in their design. The former solution adjusts a network parameter as the duty cycle while the latter employs the maximum energy consumed by an agent to budget the amounts of energy over a certain period.

Adaptive algorithms address the spatio-temporal variation of ambient energy sources by also optimising data sampling and routing in order to deliver effective power management. Efficient utilisation of energy scavenging by optimal energy allocation and packet routing decision is proposed in several works [5, 6]. Meanwhile, the adjustment of parameters as the sampling rate is preferred in other solutions with the same goal of modelling energy-neutral systems [7, 8]. While these approaches work well with the expected dynamic energy harvesting, they have the common feature that their performance is limited to the boundary of one network domain. Therefore, if the entire network is unable to harvest energy (due to ambient conditions or obstacles in the environment), no solution is enough. At the same time, the adaptive algorithms dynamically manage a node's operation to throttle its activity when energy supply is scarce and increase it during periods of high availability. Thus, these techniques may incur in the collection of undesirable data or in the loss of collectable information.

The specific scenario of cooperation studied here is a novel setup in the area of EHWSNs, where nodes require to optimise their use of harvested energy in order to fulfil as much as possible their energy requirements by collaboration. From this perspective, the authors of [9] propose an approach for efficient energy management that facilitates the energy exchange between homes equipped with renewable energy technology and storage in remote communities. Their linear programming framework for connecting agents is used as a reference point in our work.

2.2. Heuristic negotiation strategies for multi-issue negotiation

In this work, we aim to extend the standard form of optimisation to leverage an area beyond a network's boundary. To achieve this, the cooperation problem has been studied extensively by researchers in the fields of WSN and Game Theory [10, 11, 12]. In these works, a network is supposed to be rational and self-interested. Thus, there is a risk that each network aims to maximise its own benefit by utilising the other's network services but minimising its own cooperative effort. Then the main focus of the proposed research is to study the conflict situations that arise and look for possible equilibria under different network conditions. By modelling the problem of cooperation as games, the behaviour of each network and actions can be analysed to evaluate the strategic interactions and the set of possible outcomes. Despite the relevance of game theory, this approach in practice is usually highly complex and inefficient to implement [13]. The most obvious drawbacks are:

- An unbiased trustable mediator that acts to find the agreement towards the Pareto-optimal line using complete information of the players is implicit.
- The computational complexity of this search increases significantly as the number of nodes involved grows. The set of available actions for each node needs to be fully defined as well as the possible states the system can reach.

In the absence of a central device equipped with powerful computing, a WSN would necessitate nodes making a significant effort to calculate and store not only all their possible actions at each decision point but also the ones corresponding to their counterpart. Thus, the use of game theory

may demand storage and computational capacities that are not held in this domain. Similarly, the assumption of complete information is not accurate in opportunistic encounters between nodes.

To address the limitations of the game-theoretic analysis, heuristic methods are used to provide a reasoning mechanism that arises much less complexity. In this subsection, we focus on existing heuristics for automated multi-issue negotiation under incomplete information. Previous work is presented by Faratin et al. [14]. Such heuristics have been widely used in several areas and complement multiple frameworks for multi-issue negotiation [15, 16].

Work in [17, 18] proposes an offer generation technique for automated multi-issue negotiation with no information about the opponent's utility function using an alternating projection strategy. In this regard, several works have employed this strategy to develop their offer generation mechanism [19, 20]. The goal of this approach is to design strategies for computing offers when there is no available information about the participants and lead the negotiation process to an acceptable agreement for all the agents involved.

Faratin et al. [21] introduced the idea of choosing an offer similar to the opponent's preferences based on the existence of a fuzzy similarity function. However, the solution requires a similarity function that is defined for every issue of the negotiation. Such a requirement makes the mechanism domain-dependent and useful only with additive scoring functions. In contrast, works in [22, 23, 24] design a strategy, where an agent calculates offers close to the opponent's bids that match its own utility level without any additional similarity-based mechanism or information on the opponent's model. The orthogonal strategy proposed in [22] is applied in our domain of energy negotiation to address the possibility of finding an acceptable deal in finite time convergence over an interdependent multi-issue negotiation. The orthogonal strategy has proved to approximate Pareto-efficient bargaining solutions.

2.3. Reinforcement learning in the WSN domain

Reinforcement learning is particularly suitable for dynamic environments such as WSNs, where the state of current conditions can vary over time. By performing actions and adapting future decision-making based on the observed consequences of those actions, an agent can learn an optimal policy to optimise a particular objective.

In [25], the authors apply the Least-Squares Policy Iteration (LSPI) algorithm to manage cross-network optimisation problems. LSPI is used as the reasoning method to find the optimal set of network services in each WSN node. A central and powerful negotiation engine is assumed to continuously collect information about the system measurements and environmental states. The engine computes the configurations for each participating network so that the activation of the corresponding services positively influences the performance of each network. Along with the aid of a centralised decision maker, their paradigm referred to as Symbiotic Networking contemplates the integration of different networks from their design and not opportunistically.

Most of the proposed reinforcement learning based approaches solely focus on solving the WSN routing problem. Solving such a problem is found to be NP-hard. However, similarly to our case, the application of reinforcement learning in routing seeks to predict the full path quality between nodes by reducing the complexity considering only neighbouring nodes' information [26]. Each node independently performs the routing procedures to decide the minimum cost path, which leads to a near-optimal routing decision with a very low computational complexity.

The use of LSPI is reproduced in [27] to enable a node to learn an optimal routing scheme with multiple optimisation goals among the maximisation of its network lifetime. Similarly, work in [28] proposes a routing policy conditioned by the message importance that includes the selection of paths with the highest delivery rate learned over the previous routing experiences. The underlying approach is based on Q-Learning. Although the space of options is simplified in the routing domain, these techniques need to consider the set of state-action pairs to find an optimal action-selection policy. As a result, the computational complexity of the algorithms increases as the dimensionality of the problem proportional to the state representation grows.

The drawbacks of high computation complexity and large memory requirement in comparison to more sophisticated learning algorithms are reduced with MAB learning. The space of options in MABs is characterised only by the set of the agent's actions. The MAB model is commonly used in the online learning literature for solving resource allocation problems. One solution in the context of WSNs is multi-armed bandit based energy management (MAB/EM) [29]. MAB/EM is a power management technique to adapt the operation of nodes to the environmental changes while maximising the total amount of information collected over a period of time. In MAB/EM, the energy of an agent is intelligently allocated to the tasks of sampling, reception, and transmission of data, as the agent learns which combinations optimise its performance in long-term information collection. The allocation problem is also addressed in [30] by using MAB algorithms to make efficient use of the radio spectrum and avoid collision between cognitive nodes. In the model, the nodes are not aware of the medium conditions, and they have to estimate the channel's availability by exploring and learning.

Other MAB techniques are found in the context of long-range IoT networks. Radio connectivity for short range applications (e.g. WiFi, Zigbee and Bluetooth) is not suitable for covering very large scenarios [31, 32]. Low-Power Wide Area Network (LPWAN) technologies are a good choice for these emerging applications and opportunities. In this respect, the LoRaWAN (Long Range Wide Area Network) specification is a promising solution due to its Adaptive Data Rate (ADR) scheme. Recently, a MAB based approach has been proposed to optimise the performance of LoRAWANs as an alternative to the standard ADR algorithm [33]. In this work, the authors use MAB algorithms to manage the trade-

off between energy consumption and packet loss by adapting communication parameters such as spreading factor and transmission power. The comparison between ADR and multiple MAB algorithms shows a higher efficiency of the proposed MAB approach in terms of energy consumption, packet loss, and cumulative cost. Specifically, the Switching Thompson Sampling with Bayesian Aggregation policy outperforms the rest of the techniques for the studied non-stationary stochastic environment.

With a growing number of interconnected devices in IoT, dynamic spectrum access can mitigate the expected connectivity demand. Work in [34] applies MAB to support decentralised decision-making of RF channels allocation in a non-stationary environment. Two approaches are selected to compare the successful communication rate, the naive or random selection approach, and the stochastic MAB-learning based approach. The latter implemented using UCB and Thompson Sampling, empirically determines an efficiency proportional to the number of intelligent dynamic nodes. The stochastic bandit algorithms have near-optimal performance even when the number of smart objects in the network increases, which suggests that the learning methods are applicable in this non-stochastic setting. Channel-hopping is a spectrum access technique to improve the reliability of wireless networks. In the case of failure, retransmission occurs using a different channel. In this regard, the work in [35] introduces the use of MAB algorithms to enhance the spectrum utilisation by learning the channels that achieve the highest delivery ratio.

In a different context, the resource allocation problem is studied under an adversarial bandit model in which the reward or the environment dynamics cannot be attributed to a distribution function. From this perspective, [36] formulates a joint power control and channel selection strategies in infrastructure-less wireless networks. These strategies can be used to achieve the equilibrium for efficient resource management and interference mitigation among selfish transmitters. In this paper, we also pose our partner selection problem as an adversarial MAB problem, in order to optimise the energy management of the network for long-term periods. We evaluate the performance of several algorithms on partner selection through practical scenarios in WSNs to have an accurate online estimation method of whether a particular policy will work well in practice. In this regard, there are no prior results for the reward maximisation on partner selection.

2.4. Partner selection in automated negotiation

The problem of partner selection has been studied from different perspectives. In [37], a motivation-based mechanism maps goals and issues to motivations. The mechanism uses the history of candidates' performance to select those that have the most beneficial effects in terms of current motivational needs. In other settings, the negotiation outcome and its equilibrium are analysed in terms of the amount of information that is known about the opponent's parameters [38]. The results reported are useful for decision making in situations where an agent has the option to select a partner on the basis of the information state about its opponents.

In [39] the authors propose a framework for automated negotiation based on negotiation profiles. Each agent gathers information during the negotiations and stores it in the associated profile: the preference profile keeps the agent negotiation strategy, the partner cooperation profile records the agent interaction with the other agents in the environment, and the group-of-partners negotiation profile stores the profiles of several negotiation partners. The agent is then able to construct a set of rules to anticipate both the outcome and the best potential partner with which to start a negotiation. A central facilitator is responsible for registering new agents and informing others about it.

The problem of partner selection in [40] is analysed using a possibilistic case-based decision model. Their solution provides the decision theoretical basis to predict the possibility of successful negotiation with other agents using small historical data about past negotiation behaviour and the derived qualitative expected utility for a specific situation. Accordingly, they keep a record of past negotiations to model the negotiation behaviour of the opponents and be able to predict it in the future.

As shown by previous research, the record of past negotiations or related information is essential for choosing the negotiation partner among a set of candidates. This makes sense in devices equipped with advanced processors and large memory capacity. In fact, the design of automated negotiation is highly sensitive to the domain in which the interactions take place. In our domain, the networks are resource constrained systems that discover each other opportunistically and have no information about their neighbours. Moreover, the widespread use of WSNs predicted in the future and the increasing likelihood of different WSNs deployed in the same place demand a proper policy to aid an agent on the decision-making process of the most prospective partner. In this environment, the most promising partner is evaluated in terms of agreements on energy cooperation, where the position of the nodes and the orientation of their energy sources strongly impact the energy harvested. That is, even if two nodes are geographically close, their harvesting rates may vary significantly.

3. System models and problem formulation

The basic idea of OEN between multi-domain and co-located sensor networks is to define an agreement on the energy flow that handles the spatio-temporal variation of the participants' energy sources and satisfy as much as possible their load through a collaborative effort. Thus, in multi-agent negotiation, the main goal for an agent is to select the best potential partner that would maximise its energy allocation.

This section defines the network model and the assumptions considered in our work. The energy management model, the strategic negotiation model and the partner selection problem for efficient energy allocation in the long term

are also presented.

3.1. Network model

We consider a set N of m energy-harvesting wireless sensor networks $N = \{N_1, N_2, ..., N_m\}$ that are under the administration of distinct authorities and deployed in the same area. Each network involved N_i, $1 \leq i \leq m$ has distinctive characteristics and is formed by a set of unique sensor nodes $N_i = \{1, ..., j, ..., |N_i|\}$ and a sink.

Our scenario consists of general WSN applications that periodically collect data from the sensor nodes and report these measurements to the sink, using multiple hops to traffic the packets. For the sake of simplicity, we assume that time is divided into discrete time slots $T = \{1, ..., n\}$ of equal duration L, and each time slot t is long enough to deliver all packets to the collector and take a decision about inter-network cooperation.

We examine each network N_i as a cooperative multi-agent system. Then, each node j in N_i is controlled by an agent. The identity of an agent is indexed by i, j, $1 \leq i \leq m$, $j \in \mathbb{N}$. The agent has complete knowledge of all the relevant node's information, such as its neighbours, its energy availability at each time slot t, including the availability in the future, its load, its battery capacity and residual energy. In general, a node is an autonomous agent with advanced situational awareness of itself and its local neighbours (the nodes in its own network).

Suppose that we define a neighbourhood of an agent i, j as $\Omega_{i,j}$, such that $\Omega_{i,j} \subseteq N_i$, and that the agent i, j knows about all other agents in its 1-hop neighbourhood. Thus, a neighbourhood is a subset of agents in N_i that control sensors with overlapping radio range. Opportunistic Direct Interconnection (ODI) is possible between any pair of nodes whose communication beams overlap [41]. Since the main focus of the paper is on the partner selection of each agent i, j, we make the following assumptions about the neighbourhoods, the pre-negotiation communication phase and the transmission properties:

(a) Neighbourhoods do not necessarily have the same number of members, and each agent i, j belongs to only one neighbourhood $\Omega_{i,j}$.

(b) OEN is proactive: each agent i, j periodically broadcasts HELLO messages that contain its energy status and the list of its current neighbours along with their status. In this way, each agent i, j can maintain a map of energy conditions across its neighbourhood $\Omega_{i,j}$. To eliminate OEN overhead, HELLO information is introduced onto the broadcast updates required by the used routing protocol.

(c) If multiple agents discover a lack of energy at the same time in the same neighbourhood, we assume that agents are assigned a priority level and rotate with time. The assignment mechanism is out of the scope of this work.

(d) No packet loss occurs during cross-network communication. This is relevant for the delivery performance of offers during the negotiation process. It is a valid assumption since no loss is observed under the introduction of ODI architectures [41].

(e) Each agent addresses the communication cost of the iterative exchange of offers before the negotiation. We assume that this cost is negligible compared to the energy aimed by a node to win after negotiation. This assumption is reasonable in negotiations with pre-established short deadlines [42]. Moreover, the experimental results found in [43] shows that the energy cost to maintain ODI functionality is also insignificant.

The nodes in N_i usually operate unattended in a collaborative manner to perform some tasks. Such tasks include sampling, reception, processing and transmission. Although the execution of these tasks consumes a measurable amount of energy, we ignore the power used up in processing and sampling since the communication energy for reception and transmission is a dominant factor in most sensor platforms. Given this, the total energy consumed by an agent is in terms of its radio transceiver's duty cycle. Furthermore, each agent i, j consists of an energy harvester unit and a rechargeable battery. The energy management model used to derive an agent's energy profile is described in the next subsection.

3.2. Energy management model

In this work, we refer to a simplified model of average power consumption as it is used in [44]. Let $E^c_{i,j}(t)$ denote the energy consumed by radio communication of agent i, j in time slot t. Then we have:

$$E^c_{i,j}(t) = V \cdot \Big[D \cdot I^{active} + (1 - D) \cdot I^{sleep} \Big] \cdot L \tag{1}$$

The maximum energy an agent can spend at each t for some duration L depends on the duty cycle D, supplied voltage V, active mode current I^{active}, and sleep mode current I^{sleep}. D is set by the node's application, while I^{active},I^{sleep} and V can be known in advance using datasheet information.

The energy consumption along with the agent's energy availability is used to compute the energy allocation of a WSN. The energy management model is built on the proposals made by [3] and [9]. We take the power management characterisation from [3] and model a linear function as [9] to design the utility function of each agent i, j.

Let $E^{hrv}_{i,j}(t)$ and $E^c_{i,j}(t)$ denote the energy profile variables for each time period, where the former is the amount of energy that can be generated by the harvesting source in slot t. We consider that the agent can forecast $E^{hrv}_{i,j}(t)$ from historical information with high accuracy. Let $B_{i,j}(t)$ denote the residual energy of the battery at the beginning of slot t in agent i, j. Then, the battery energy left after the last slot of the energy harvesting period is defined as $B_{i,j}(n + 1)$. The cycle of the battery is represented by $B_{i,j}(n + 1) = B_{i,j}(1)$. The battery is characterised by a limited energy capacity $B^{max}_{i,j}$ and charging efficiency e. When $E^{hrv}_{i,j}(t)$ is lower than $E^c_{i,j}(t)$, some of the energy used by the sensor node is discharged from the battery. We use $d(t)$ to represent this energy amount. When $E^{hrv}_{i,j}(t)$ is higher than $E^c_{i,j}(t)$, all the

energy used in the node is provided by the harvested source and the battery is charged with the excess as required, up to its maximum capacity $B_{i,j}^{max}$. We use $c(t)$ to denote this energy amount. Based on this, the energy used from the battery in any slot t can be calculated as:

$$B_{i,j}(t) - B_{i,j}(t+1) = d(t) - e \cdot c(t) \tag{2}$$

An agent can use its battery to save and spend energy over the entire period of n slots, which allows the agent to compute an energy allocation for each t. The energy allocation at time slot t is denoted as $E_{i,j}^{alloc}(t)$. The energy that the agent is unable to use or store at time slot t represented by $w_{i,j}(t)$ is wasted. An opportunistic energy negotiation process is initiated when an agent's estimated energy level is not enough to maintain the next period. Thus, the initial battery status $B_{i,j}(1)$ is equal to $e \cdot b$ where b is the energy level at $t = 1$.

During the negotiation, agent i, j considers the amount of energy to receive/give from the cooperation effort at each time t, which is defined by $o = (o(1), ..., o(n)) : o \in \mathbb{R}_{+}^{n}$. o represents the offer of energy at each time slot, i.e., the issues of our negotiation domain. We call these offers *energy flow offers*. A valid proposal must include the energy values along with the corresponding sign (if positive, the amount is an offer of energy from the agent to its opponent, otherwise, it represents the energy to be received from the opponent.) for the predetermined time of cooperation, e.g. If networks expect to cooperate for 6 hours, then the energy flow must include 6 values when L=1 h.

Since the current battery status only depends on the amount of energy harvested and consumed during previous slots, as represented in equation (2), the energy allocation problem can be formulated as a linear program. The objective function of this program is to maximise the agent's utility. The agent's utility corresponds to the total energy allocated to power a load over the time interval [1, n], given as:

$$u_{i,j} = \sum_{t=1}^{n} E_{i,j}^{alloc}(t), \tag{3}$$

where the utility of agent i, j is represented by $u_{i,j}$ and describes the total amount of energy consumption that can be satisfied at period T.

An optimal allocation of energy can thus be obtained by solving the following linear programming problem:

$$Objective \quad \text{maximise} \quad u_{i,j} \tag{4}$$

Subjected to the following constraints:

- The allocated energy at time slot t, $E_{i,j}^{alloc}(t)$, is defined by the harvested energy, the charged and discharged energy from the battery, the energy flow offer and waste:

$$E_{i,j}^{alloc}(t) = E_{i,j}^{hrv}(t) - c(t) + d(t) + o(t) \tag{5}$$

- The following represents the energy balancing condition, which determines that the allocated energy $E_{i,j}^{alloc}(t)$ must not exceed the maximum amount of energy $E_{i,j}^{c}(t)$ that an agent can consume at slot t:

$$E_{i,j}^{alloc}(t) \leq E_{i,j}^{c}(t) \tag{6}$$

- The energy used from the battery at any time t depends on the discharged $d(t)$ and charged $c(t)$ energy plus its efficiency e:

$$B_{i,j}(t) - B_{i,j}(t+1) = d(t) - e \cdot c(t) \tag{7}$$

- The battery level at time slot $t = 1$ is equal to an initial residual energy b:

$$B_{i,j}(1) = e \cdot b \tag{8}$$

- The cycle of the battery is represented as:

$$B_{i,j}(n+1) = B_{i,j}(1) \tag{9}$$

- The energy stored into the battery at each time t, $c(t)$, cannot be negative and must not exceed the maximum battery capacity:

$$0 \leq c(t) \leq B_{i,j}^{max} \tag{10}$$

- The energy drawn from the battery at each time t, $d(t)$, when $E_{i,j}^{hrv}(t) < E_{i,j}^{c}(t)$ starts from $E_{i,j}^{c}(t) - E_{i,j}^{hrv}(t)$. This amount must also not exceed the residual energy of the battery:

$$E_{i,j}^{c}(t) - E_{i,j}^{hrv}(t) \leq d(t) \leq B_{i,j}(t) \tag{11}$$

- At each time t, the battery must not store more energy than its capacity, also it cannot have negative values:

$$0 \leq B_{i,j}(t) \leq B_{i,j}^{max} \tag{12}$$

- Any wasted energy in t is positive and cannot exceed the energy harvested $E_{i,j}^{hrv}(t)$:

$$0 \leq w_{i,j}(t) \leq E_{i,j}^{hrv}(t) \tag{13}$$

When o is null, given the node's detailed energy profile describing its maximum load $E_{i,j}^{c}$ and energy harvested $E_{i,j}^{hrv}$ on interval [1, n], initial battery status $B_{i,j}(1)$, battery efficiency e and battery capacity $B_{k,i}^{max}$, an agent can easily compute the node's utility over T around $E_{i,j}^{alloc}$, c, d, $B_{i,j}$ and $w_{i,j}$ as outlined in Algorithm 1. This energy management method forms the basic scheme for energy allocation and gives the sensor network self-organising ability to anticipate an insufficient energy provision.

The algorithm meets the conditions listed in Subsection 3.2 to optimise the objective function of energy allocation. Our work focuses on simple solutions for resource constrained sensor nodes. Algorithm 1 defines the energy allocation steps a node follows to compute an energy allocation scheme without adapting its network parameters. The

Algorithm 1: Agent's utility without OEN

Input : $E^{hrv}_{i,j} \in \mathbb{R}^n_+$, $E^c_{i,j} \in \mathbb{R}^n_+$, $B_{i,j}(1) \in \mathbb{R}^+$, $e \in [0,1]$, $B^{max}_{i,j} \in \mathbb{R}^+$, $n \in \mathbb{Z}^+$;

Output: $E^{alloc}_{i,j} \in \mathbb{R}^n_+$, $c \in \mathbb{R}^n_+$, $d \in \mathbb{R}^n_+$, $B_{i,j} \in \mathbb{R}^n_+$, $w_{i,j} \in \mathbb{R}^n_+$

1 Initialisation: $E^{alloc}_{i,j}(t) = 0, c(t) = 0, d(t) = 0, B_{i,j}(t) = 0, w_{i,j}(t) = 0$ for $t = 1, 2, ..., n$;
2 first = True ;
3 **for** $t \leftarrow 1$ **to** $n-1$ **do**
4 **if** *first* **then**
5 $B_{i,j}(t) = B_{i,j}(1)$;
6 first = False;
7 **end**
8 **if** $E^{hrv}_{i,j}(t) \geq E^c_{i,j}(t)$ **then**
9 $E^{alloc}_{i,j}(t) = E^c_{i,j}(t)$;
10 **if** $E^{hrv}_{i,j}(t) - E^c_{i,j}(t) > \frac{1}{e} \cdot (B^{max}_{i,j} - B_{i,j}(t))$ **then**
11 $c(t) = \frac{1}{e} \cdot (B^{max}_{i,j} - B_{i,j}(t))$;
12 $w_{i,j}(t) = E^{hrv}_{i,j}(t) - E^c_{i,j}(t) - c(t)$;
13 **else**
14 $c(t) = E^{hrv}_{i,j}(t) - E^c_{i,j}(t)$;
15 **end**
16 **else**
17 **if** $E^c_{i,j}(t) > E^{hrv}_{i,j}(t) + B_{i,j}(t)$ **then**
18 $E^{alloc}_{i,j}(t) = E^{hrv}_{i,j}(t) + B_{i,j}(t)$;
19 $d(t) = B_{i,j}(t)$;
20 **else**
21 $E^{alloc}_{i,j}(t) = E^c_{i,j}(t)$;
22 $d(t) = E^c_{i,j}(t) - E^{hrv}_{i,j}(t)$;
23 **end**
24 **end**
25 $B_{i,j}(t+1) = B_{i,j}(t) - d(t) + c(t)$;
26 **end**
27 **if** $E^{hrv}_{i,j}(n) \geq E^c_{i,j}(n)$ **then**
28 $E^{alloc}_{i,j}(n) = E^c_{i,j}(n)$;
29 **if** $E^{hrv}_{i,j}(n) - E^c_{i,j}(n) > B_{i,j}(1)$ **then**
30 $c(n) = B_{i,j}(1)$;
31 $w_{i,j}(n) = E^{hrv}_{i,j}(n) - E^c_{i,j}(n) - c(n)$;
32 **else**
33 $c(n) = E^{hrv}_{i,j}(n) - E^c_{i,j}(n)$;
34 **end**
35 **else**
36 **if** $E^c_{i,j}(n) > E^{hrv}_{i,j}(n) + B_{i,j}(n)$ **then**
37 $E^{alloc}_{i,j}(n) = E^{hrv}_{i,j}(n) + B_{i,j}(n)$;
38 $d(t) = B_{i,j}(1)$;
39 **else**
40 $E^{alloc}_{i,j}(n) = E^c_{i,j}(n)$;
41 $d(t) = E^c_{i,j}(n) - E^{hrv}_{i,j}(n)$;
42 **end**
43 **end**

algorithm is optimal since it uses all the energy harvested to power the agent's load. The initial battery status is used as a sign of insufficient energy allocation. Such an event can occur when the energy availability is affected either due to obstructions of power source, damaged batteries or inefficient energy sources.

In every time slot t, the algorithm evaluates two cases depending on the data of E^{hrv} and E^c: when there is enough energy harvested to complete a load (Step 8) and the second case when the energy availability is attempt to be supplied with the help of the battery (Step 16). E^{alloc}, B, c, d and w are derived from Subsection 3.2 given the data of E^{hrv}, E^c, $B(1)$, e and B^{max}. Then, the problem can be solved for any t if $B(t-1)$ is known.

In more detail, when there is excess energy and it goes above the battery capacity (Step 10), the battery is charged with the excess as required taking into account its greatest capacity and the rest is discarded. Otherwise (Step 13), the battery is only charged with the excess. Step 16 depicts the scenario when there is not enough ambient energy to power a load. There are two cases to evaluate in this statement: when the battery cannot supply the missing energy (Step 17), and the opposite (Step 20). In every case, the values for energy allocation and discharge are depicted.

At the end of the algorithm run, the resulting energy allocation scheme describes the situations that can be considered by agent i, j to decide if an OEN with a co-located network must be performed, i.e when an agent can not harvest enough energy for its consumption, and the difference can not be covered with the residual capacity of its battery. The statement in Step 27 describes the conditions for the last slot in the window and cycle of the battery.

Algorithm 1 can be used to automatically alert the agent if a deficient energy allocation scheme is envisaged. When o is not null and we include offers, a centralised mechanism to compute the optimal energy flow that benefits both agents is the Nash Bargaining Solution (NBS) [45]. However, NBS requires complete information, a trusted neutral third party, and high computation capabilities since the set of all possible agreements is exponential in the number of time slots. NBS is used as a benchmark for the performance evaluation of our approach. The next subsection describes the negotiation model used for the bargaining process.

3.3. Energy negotiation model

We focus on bilateral negotiations, i.e. negotiations between two neighbouring rechargeable agents that belong to distinct WSNs. In bilateral negotiation, negotiation strategies are critical. This subsection describes the protocol and strategies used following a generic framework for automated multi-issue negotiation [23]. The model is empirically evaluated at the end of the subsection to show how effective the mechanism is in comparison to a cooperative game theoretic approach, based on the Nash bargaining solution.

In a bilateral negotiation, both agents are willing to cooperate but have conflicting interests regarding their preferences (in this domain due to distinct batteries, power con-

sumption and energy harvesting profiles). To that end, agents have to negotiate and determine the most beneficial setup before cooperation. Specifically, Rubinstein's alternating-offers protocol [46] is adopted for the interaction between agents. At each round, an agent can either accept an offer from the opponent, reject it, propose counter-offers, or opt out of the negotiation, usually once the negotiation deadline is reached.

In OEN, the negotiation proceeds in a sequence of rounds $R = \{1, 2, 3, ..., r_{max}\}$ for a predefined short-term deadline r_{max}. As defined in 3.2, $o = (o(1), ..., o(n))$ represents the vector of issues to be negotiated in each negotiation round r. The agents in our domain only propose one offer in each round. Thus, $o^r_{1,1\to2,1}$ is a vector of values proposed by agent 1_1 to agent 2_1 at round r, where $o^r_{1,1\to2,1}(t)$ is the value of energy offered from 1_1 to 2_1 for time slot t. The issues in this domain maintain interdependencies between each other due to the use of the battery. For a time slot t, the energy flow (energy going out/into the agent) depends on how much energy an agent harvest or how much energy had been stored/withdrawn in previous time slots. We assume the negotiation context (issues, deadline and initial negotiating agent) is known by both agents beforehand, and it remains unchanged during the whole encounter.

In our work, agents adopt time-dependent strategies [14] to determine the amount of concession required for each offer. At the first round, the agents propose deals that give the highest utility to themselves. Afterwards, different agents may have different attitudes towards deadlines. Rounds conduct the values of the negotiation issues, the more rounds has passed the more pressure is induced and faster concessions are possible. The agent can adopt two behaviours: it may be impatient to reach a deal, so it concedes quickly and the offer rapidly changes to the reservation value (Conceder agent), or it may adopt a tougher strategy and maintain its initial proposal until it almost approaches the deadline (Boulware agent). In our experiments, the following time-dependent function is employed as the concession strategy to model the target utility value (the amount of energy allocation desired for the period T) of an agent i, j at each round r of the negotiation:

$$u^r_{i,j} = min_{i,j} + (1 - \alpha^r) \cdot (\sum_{t=1}^{n} E^c_{i,j}(t) - min_{i,j}) \tag{14}$$

where $min_{i,j}$ denotes the reserved value of agent i, j over T i.e., the minimal amount of energy an agent i, j can allocate for its consumption when o is null, found by Algorithm 1. The sum $\sum_{t=1}^{n} E^c_{i,j}(t)$ is the maximum amount of energy an agent can allocate to power its load over T. Then, the target utility at each round is within the range $[min_{i,j}, \sum_{t=1}^{n} E^c_{i,j}(t)]$. Function α^r is parameterised by the scaled

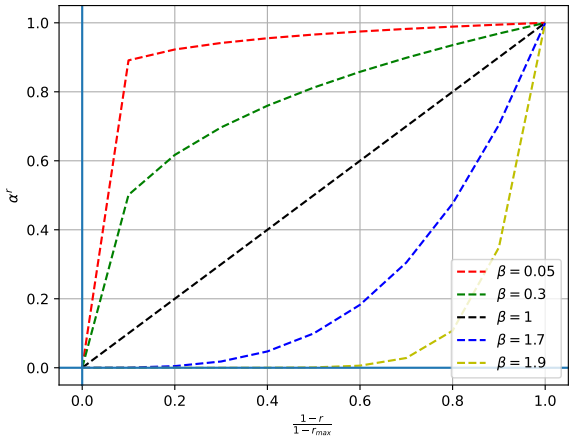

Figure 3: Polynomial function for the computation of $u^r_{i,j}$. Negotiation round is presented as relative to r_{max}.

round and concession rate β as follows:

$$\alpha^r = \begin{cases} \left(\frac{1-r}{1-r_{max}}\right)^{\beta}, & \text{if } \beta < 1 \\ \left(\frac{1-r}{1-r_{max}}\right)^{\frac{1}{2-\beta}}, & \text{if } 1 \leq \beta < 2. \end{cases} \tag{15}$$

Following this strategy, the shape of the concession curve represents a human's negotiation behaviour. If $\beta < 1$, agent i, j adopts a Conceder behaviour; if $1 < \beta < 2$, the agent uses a Boulware tactic (see Fig. 3).

Given an agent's utility function at round r, an agent can define its response for an opponent's offer. Thus, if agent 1_1 receives an offer $o^r_{2,1\to1,1}$ from agent 2_1 at round $r < r'$, the interpretation of agent 1_1 defined as H at round r' for the opponent's offer is given by:

$$H^{r'}_{1,1}(o^r_{2,1\to1,1}) = \begin{cases} accept, & \text{if } u^r_{1,1}(o^r_{2,1\to1,1}) \geq u^{r'}_{1,1}(o^{r'}_{1,1\to2,1}) \\ reject, & \text{otherwise.} \end{cases} \tag{16}$$

If the offer is rejected, the agent in turn proposes a new agreement, which again the opponent may accept or reject in the next round. The negotiation will continue until an offer is accepted, a final negotiation round is reached, or the process is terminated by any of the participants (ending it with no deal possible).

Agents' strategy of generating offers is implemented using the orthogonal strategy [22]. The main idea behind the orthogonal strategy is to always select the point which is the closest (measured in the Euclidean distance) to its opponent's last offer on its indifference curve (i.e., the points that give the same utility for the agent). Let $o^{r-1}_{2,1\to1,1}$ be the last offer from agent 2_1 to agent 1_1 at round $r - 1$. If agent 1_1 needs to generate a counter proposal that lies on the indifference curve C according to its target utility $u^r_{1,1}$,

then agent 1_1's offer at round r with the shortest distance to $o^{r-1}_{2,1\to1,1}$ can be calculated by:

$$o^{r}_{1,1\to2,1} = \arg\min_{o\in C} \|o - o^{r-1}_{2,1\to1,1}\| \tag{17}$$

where $\|.\|$ denotes Euclidean distance and the offer is subjected to the constraints listed in 3.2. Since energy is logically transferred between networks by accepting energy-consuming tasks like data processing or packet forwarding, a change is required in constraint (10) when there are offers involved:

$$0 \leq c(t) \leq E^{hrv}_{k,i}(t) \tag{18}$$

The result is that the battery will be charged immediately with the energy harvested by the agent while the energy supply received from the opponent's offer will be used to satisfy the agent's load.

In summary, the mechanism presented in this subsection allows an agent to represent its preferences and determine the desired utility level to generate a counter-offer accordingly. Our system makes use of the described negotiation model in order to fulfil the network objective of long-term energy allocation.

3.3.1. Performance Evaluation of OEN

These experiments were conducted with different agent's profiles and negotiation parameters. Two distinct types of ambient energy sources (solar and wind) were considered to evaluate our approach experimentally: Agent *a* controls a sensor node with a solar panel while agent *b* manipulates a sensor node with a wind turbine.

We acquire hourly wind speed collected at Weather Underground [47] and solar radiation from PVGIS [48] for a period of one year (2017) corresponding to the area of Southampton to compute the energy generated by solar panels and wind turbines. The computation of solar energy considers a panel of dimension 3.3 cm × 6.35 cm with a maximum efficiency of 10%. Then, the estimated hourly power output is proportional to the solar radiation obtained from PVGIS, the panel dimension, and its efficiency [49]. For the wind source, the power is calculated using the wind speed and swept area of 5 cm × 5 cm as in [50]. Then, the values are scaled to get the hourly power output of a highly efficient micro-turbine.

In order to obtain the following results and compare OEN to NBS, a feasible space of energy flow agreements must exist between the agents. Thus, energy harvesting values from the same day, same period, are selected to meet the requirement of feasible intersection points. Regarding the power consumption profile, agent *a* and *b* are modelled with a supply voltage of 3 V, sleep current of 5 µA, active current of 20 mA and duty cycle drawn from the discrete uniform distribution on the interval [1%, 5%]. These values are used to determine the regular energy consumed by each agent at every hour. The parameter values for the agents' storage are a maximum capacity of 600 mAh and efficiency of 70%, which is typical for NiMH batteries.

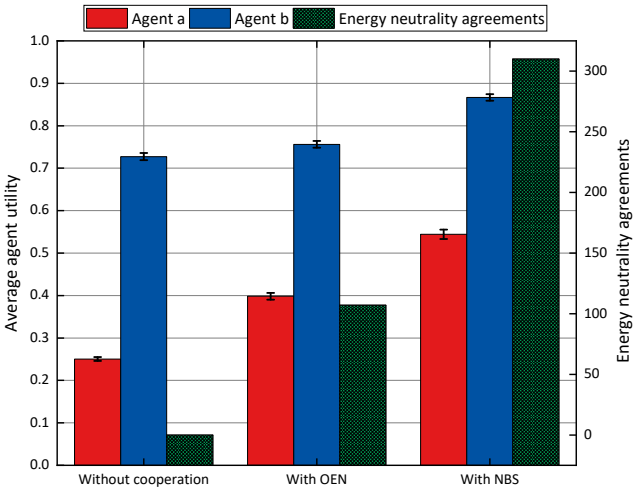

Figure 4: Average utility of agents a and b in the following situations: without cooperation between them, with negotiation using OEN, and the theoretical optimum NBS. The error bars denote the standard error to the mean. The graph also displays the comparison between energy neutrality agreements found with each solution.

In terms of the negotiation parameters, the concession rates for each agent and the number of rounds at each simulation are chosen randomly from 0.5 to 1.5 and 5 to 15, respectively. The number of issues is fixed at 6, which means that at every encounter, the agents negotiate for a cooperation period of 6 hours.

With the aim to optimise its power management, the first step for each agent is to identify its own efficiency. In OEN, it corresponds to the energy allocation scheme that an agent can employ to power its load. Thus, each agent first invokes the optimal energy allocation algorithm (Algorithm 1) to make optimal use of its harvested energy. This power management technique is tested during every simulation. Such algorithm enables self-organised agents that can anticipate insufficient energy allocation schemes and the opportunity to start an OEN. The main goal of the simulations is to compare the performance of the agents in the following situations: without cooperation, with optimal energy allocation after negotiation incorporating the OEN model, and finally using the Nash bargaining solution.

Fig. 4 shows the average utility value obtained over 1000 bilateral negotiations for agents: *a* and *b*. As observed, on average, the energy harvested from solar and wind power sources is insufficient to satisfy the total load of each agent independently. On the other hand, the energy allocation with OEN is improved on average up to 59% and 4% as a result of agreements obtained by the agents on energy cooperation, in agent *a* and *b*, respectively. This amount of energy can be up to 50% of the energy that is reallocated using the Nash bargaining solution. From these cases, 107 energy flow agreements meet the energy-neutrality condition when applying

OEN, while 310 are achieved by the centralised solution. The percentage of energy neutrality deals with OEN are 35% of the optimal solution. Although energy neutrality is not imposed in the offers proposed using OEN, the mechanism is capable to reach this most desirable outcome. Moreover, the introduction of weather data positively correlated in our simulations can make cooperation even more successful and increase the possibility of reaching better results. The combination of sun and wind energy may ensure every agent to exploit its source to the fullest.

The performance of OEN can be increased by exploiting co-located nodes that allow better energy allocation agreements to the agent. Therefore, an agent should be able to anticipate the best potential partner with which to start the negotiation process at every opportunistic encounter. Instead of choosing randomly an opponent between the co-located nodes, an agent should implement a better decision making policy for the selection of the most suitable opponent.

3.4. Partner selection problem for long-term energy allocation

Given the agent utility function defined in equation (3), the utility function for the whole network N_i is:

$$u_i = \sum_{j \in N_i} u_{i,j} \tag{19}$$

Thus, the global objective is to maximise the total energy allocation in the WSNs over a period of cooperation T. In other words, the network utility function is maximised when the sum of all agent's utility functions for the energy allocation is maximised, which may imply that the communication between all agents in the network is required. However, this is not the case for OEN, which considers the suboptimal approach where interactions between agents are performed only with the agents in the same neighbourhood Ω. Each agent i, j can maintain a map of energy conditions across its neighbourhood $\Omega_{i,j}$ employing broadcast messages used by the routing protocol and avoiding the possible communication overhead. Consensus via local communication with their neighbours is only required when more than one agent discover a lack of energy at the same time in the same neighbourhood. In that case, agents must decide the order in which negotiations take place.

OEN presents a different proposal from other related works, where agents adjust their duty cycle according to the energy availability in the environment and are limited by the bounds of one domain. In contrast, our model allows the agents to set a desire energy consumption for the next time slots and extend their power management strategies to an inter-network approach in order to satisfy their energy allocation through cooperation with a neighbour network. The networks that we are studying attempt to maintain an energy-neutral operation, i.e., the energy harvested is sufficient to satisfy the energy used during the same time. Energy neutrality can be supported by the following condition:

$$B_{i,j}(t) \geq E^c_{i,j}(t) \tag{20}$$

The constraint (20) enforces that the residual $B_{i,j}(t)$ at each slot must be bigger than the energy consumption $E^c_{i,j}(t)$ of any agent i, j in the network at any time slot. However, the assigned energy budget $B_{i,j}(t)$ depends on the energy availability and negotiation strategy of each negotiator involved. We then relax this requirement and measure the utility of an agent given the power management strategy of OEN. The aim is to study the effects and potential of cross-boundary energy transfer, where the most prominent outcome in this scenario is the achievement of energy neutrality for all participants. Ultimately, the goal of an agent is to decide and choose an opponent/partner so as to maximise its energy allocation in the long term.

The domain studied in our work consists of multiple overlapping networks constructed in the same area. In such a situation, there are a number of co-located agents that belong to different networks with different behaviours and network goals. The energy allocation is expected to be optimised through negotiation and cross-network optimisations. It means that at each opportunistic encounter, an agent needs to select one or more agents in the neighbour networks as the most prospective negotiation partners with whom the expectation of successful negotiation and the achievement of the best agreement are the highest.

The selection method of an appropriate partner must be able to learn the dynamism of the environment and adversarial setting introduced by the negotiation behaviour of the opponents. To solve the partner selection problem, we proposed a MAB based partner selection model for each agent within the network. In probability theory, MAB learning provides a theoretical framework for sequential learning and decision-making to address the trade-offs between exploration and exploitation under uncertainty. Unlike traditional partner selection methods, which require historical data to calculate the outcomes of negotiation and predict the possibility of successful negotiation, we use MAB to estimate the profitability of each agent and develop an online (or adaptive) scheme able to tolerate dynamically changing environments and adversarial conditions without prior knowledge.

Unlike traditional partner selection methods, which require historical data to calculate the outcomes of negotiation and predict the possibility of successful negotiation, we use MAB to estimate the profitability of each agent. Specifically, we develop an online (or adaptive) scheme able to tolerate dynamically changing environments and adversarial conditions without prior knowledge.

4. Multi-armed bandits for partner selection in WSNs

In this section, we define the K-armed bandit problem formally and show how it can model the partner selection in negotiation for an efficient long-term energy allocation in

WSNs. In doing so, we discuss existing policies and later report their comparison.

4.1. The multi-armed bandit problem

The Multi-Armed Bandit problem originally proposed by Robbins [51] refers to the gambler's dilemma. Correspondingly, the goal of a gambler is to maximise the total rewards earned through a sequence of lever pulls over a row of slot machines. Specifically, a set of K machines (arms) is available to the decision maker. Each arm has a reward associated that is independently drawn from an unknown distribution when it is pulled. At each trial, a gambler must choose which of these arms to play. To keep the terminology of MAS consistent, from here onwards the term gambler is replaced by agent, and the lever pulling action of the gambler is specified as an action of that particular agent.

Without any prior knowledge on the machines' profitability, the agent can still collect partial information while it observes the reward of each chosen arm. Such information can be used to estimate the revenue of the machines. It thus becomes a dilemma, between *exploiting* the machine that has the highest expected reward or *exploring* the set of different machines to gain more information and learn about their reward density. The fundamental challenge in bandit problems is to define the pulling strategies (also referred to as policies) for decision making in situations under uncertainty to trade-off between exploration and exploitation. A MAB learning model is particularly useful to model agents that learn a hidden reward distribution while maximising their gains.

Formally, let $Tr = \{1, 2, ..., \mathrm{Tr}\}$ be a set of sequential trials and tr denote a trial in Tr. We define the action of an agent i, j at trial tr as $a_{i,j}(tr)$, which raises the reward $r_{a_{i,j}}(tr)$. An agent's objective is to maximise the sum of its observed rewards over a sequence of decisions as follows:

$$\text{maximise} \sum_{tr=1}^{Tr} r_{a_{i,j}}(tr) \tag{21}$$

As such, it is clear that the agent has to choose a policy (i.e. a sequence of actions $a_{i,j}(1), ..., a_{i,j}(Tr)$) that maximises the total rewards earned through a sequence of trials. Then, the agent has to choose at every trial tr the best single action in order to maximize its reward in Tr trials.

The performance of the policy applied by an agent at a given trial is measured in terms of *regret*, defined as the expected loss of applying the policy with respect to the maximal expected reward by a policy assumed to be optimal. In stochastic MAB problems, this notion of expected regret is often considered. However, in our domain, a different concept of regret is incorporated, suitable for the adversarial MAB problem of our environment. This notion of regret known as *weak regret* is described in the next subsection.

The perception of optimality and bandit policies vary according to the environment. The following subsection presents a description of our practical application of MAB and the existing policies applied to our specific problem.

4.2. Multi-Armed bandits formulation for partner selection in OEN

In our domain, the environment is adversarial. Unlike classical stochastic MAB problems whereby the rewards are independently drawn following a fixed but unknown distribution, for the adversarial setting, there is no statistical assumption about the generation of rewards. Instead, the rewards are chosen by an adversary.

An adversarial MAB formulation is a natural fit for modelling our research problem, where an agent and opponent interact to solve their conflicts and the opponent is adaptive. More precisely, in this context, the outcome of a negotiation between one agent and its opponent forms the reward value that the MAB model gets by selecting a partner for opportunistic energy negotiation.

In OEN, an agent has an incentive to negotiate but may adopt different behaviours based on its preferences and observations at any time. The preferences of the networks vary according to their energy availability which is influenced by the amount of energy harvested during each time slot. The networks can then adopt a responsive attitude towards their environment using conceding strategies during the negotiation. Since agents have to negotiate with incomplete knowledge of the opponents and have no control of the environmental factors affecting the outcomes of the negotiation, it is very difficult to estimate a distribution for the rewards.

In this paper, we focus on environmental changes such as varying energy availability, which influences the different patterns of the agent's negotiation behaviour. Furthermore, the topology of the networks is also intrinsically dynamic as sensors may fail, move, or enter in sleep/active state. An agent can also reject a negotiation encounter or be added opportunistically at any time. Thus, we consider this variant of the MAB problems, where the stochastic assumption about the processes of rewards is removed and their realisation rely on the agents involved, their status, preferences and negotiation behaviours.

We consider repeated bilateral negotiation encounters over a finite number of trials Tr where three or four WSNs overlap within a geographical area. In each trial tr, there are two or three agents that belong to different networks in the immediate neighbourhood of the main agent. The main agent needs to select one opponent between these two or three agents, as the most preferred negotiation partner to reach energy cooperation agreements that maximise its energy allocation. The action is easy to identify then in our domain, for each agent i, j in a wireless sensor network N_i that needs to start an opportunistic energy negotiation, an action of agent i, j at trial tr denoted as $a_{i,j}(tr)$, corresponds to the election of a negotiation partner (e.g. $a_{1,1}(1) = \{negotiation\_partner : 2\_1\}$) among a set of K opponents. This action is constant over time since our work only contemplates bilateral negotiations ("one-to-one") as a decentralised decision-making process to not require a mediator. The negotiation also includes short-term deadlines to avoid transmission overhead.

Given this, let $r_{a_{i,j}}(tr)$ be the linear reward function of

agent i, j for each trial tr, defined as the amount of energy allocation reached on agreement at the opportunistic energy negotiation tr (Equation (3) in Subsection 3.2) with a selected partner $a_{i,j}(tr)$ from a set of K opponents, the objective of network N_i to maximise the total energy allocation over a number of trials Tr, can be formulated as follows:

$$\text{maximise} \sum_{tr=1}^{Tr} \sum_{\Omega \in N_i} \sum_{j \in \Omega} r_{a_{i,j}}(tr) \tag{22}$$

Therefore, the network objective is to maximise the sum of reward functions of all agents on each neighbourhood Ω of N_i, from the OEN encounter 1 to the trial Tr. Thus, each agent i, j's chosen action (i.e. the chosen negotiation partner) will determine the value of the global network objective.

Once the action of an agent and the reward function associated with each action are defined, the partner selection problem in OEN of each agent i, j can be reduced to a MAB problem. The agent's goal then is to efficiently maximise the expected total rewards against the adaptive environment and the adversarial opponent or equivalently, to minimise the cumulated loss over time, i.e. the energy that an agent doesn't get when it misses the chance to cooperate with the best partner.

In our setting where agents have no prior knowledge about the preferences of their opponents, and the outcomes are affected by unexpected environmental factors, the achievement of low-regret bounds (i.e. high performance) is not possible with any deterministic policy, especially for our non-oblivious adversary case. Alternatively, we assess state-of-the-art policies in an adversarial setting, which aim to minimise a regret with respect to the best-fixed strategy in hindsight, i.e., the best single action over all trials by having access to the history of negotiation's outcome against every opponent at each trial. This weak regret is common in similar situations in which it is impossible to learn the optimal (adaptive) strategy, mostly because the payoffs are adversarially decided by the opponent. Thus, although the optimal strategy cannot be learned, the best-fixed strategy in hindsight becomes feasible to analyse from the history of previous negotiations. Consequently, the cumulative expected regret over Tr for agent i, j represented by $R_{i,j}$ with respect to the optimal fixed strategy is:

$$R_{i,j} = \max_{a_{i,j} \in K} \sum_{tr=1}^{Tr} r_{a_{i,j}}(tr) - \mathbb{E}\left[\sum_{tr=1}^{Tr} r_{a_{i,j}}(tr)\right] \tag{23}$$

Where the first term describes the cumulative reward by the best-fixed strategy over Tr trials and the second part corresponds to the total expected reward achieved by the policy applied in our system.

We now describe four well-known policies for this problem, define the experiment scenarios and present the performance results. We selected *ε-greedy*, *Exponential-weight Algorithm for Exploration and Exploitation* (EXP3), an EXP3 variant: EXP3.S, and *Follow the Perturbed Leader with Uniform Exploration* (FPL-UE) as the bandit strategies for our experiments. These three bandit algorithms explicitly make use of an exploration parameter, they are widely used in the MAB literature and have proven to obtain sublinear upper regret bounds with an appropriate choice of the exploration factor.

4.2.1. ε-Greedy

A well-known and low-complexity heuristic policy for the bandit problem is the ε-greedy action selection strategy [52]. The ε-greedy strategy is sketched in Algorithm 2. The policy selects at each trial tr an action with uniform random probability for a fraction ε of the trials (exploration), and choose the best arm (exploitation) with a probability $1 - \varepsilon$ (Steps 4 and 6 respectively). The specification of the exploration factor ε is made based on the experiment, i.e. there is no standard value that fit-for-all scenarios.

Algorithm 2: Algorithm ε-greedy for each agent i, j

Input : $\varepsilon \in [0, 1]$, opponents $1, ..., K$;
Output: Negotiation partner $a_{i,j}(tr)$

1 Initialisation: $\hat{r}_k = 1, pulls_k = 0, rewards_k = 0$ for $k = 1, 2, ..., K$;
2 **for** $tr \leftarrow 1$ **to** Tr **do**
3 **if** $\sim U(0, 1) \leq \varepsilon$ **then**
4 $a_{i,j}(tr) \sim U\{1, K\}$;
5 **else**
6 $a_{i,j}(tr) = \arg\max_{k \in \{1,...,K\}} \{\hat{r}_k\}$;
7 **end**
8 Receive reward $r_{a_{i,j}}(tr)$ as $E^{alloc}_{i,j}(t)$ for all t in negotiation against selected partner $a_{i,j}(tr)$;
9 $pulls_k = pulls_k + 1$ where $k = a_{i,j}(tr)$;
10 $rewards_k = rewards_k + r_{a_{i,j}}(tr)$ where $k = a_{i,j}(tr)$;
11 $\hat{r}_k = \frac{rewards_k}{pulls_k}$ where $k = a_{i,j}(tr)$;
12 **end**

Given this, it can be seen that the estimated reward ($\hat{r}_k$) of a selected action is updated using its cumulative reward ($rewards_k$) and the number of times the action k has been executed ($pulls_k$). ε-Greedy adds some randomness when deciding between negotiation partners: instead of relying always on the best partner, it randomly explores other opponents with a probability ε.

4.2.2. FPL-UE

The policy considered here is based on the online prediction scheme Following-the-Perturbed-Leader (FPL) [53]. FPL has efficient treatment of problems with a linear cost function by following the perturbed leader. The original algorithm only works for oblivious adversaries and focuses on choosing the action of minimal cost by observing the loss incurred of each selected action. Our goal, instead, is to efficiently maximise the total rewards an agent can succes-

sively achieve against an adaptive and adversarial opponent. To address this problem, we employ a novel strategy for repeated interactions, called Follow the Perturbed Leader with Uniform Exploration (FPL-UE) [54]. In this approach, the learning algorithm proposed by Neu and Bartok [55] is extended introducing uniform random exploration for the reward maximisation scenario (Algorithm 4). Similar to ε-Greedy, the selection of a probability (ε in ε-Greedy, λ in FPL-UE) determines the exploration rate of the pulling strategy. That is, the agent will uniformly randomly choose a negotiation partner with a probability λ (Step 7) and select the partner that reaches the maximum estimated reward perturbed by the noise factor z_k (Step 5) every $1-\lambda$ of the cases. We measure the efficiency of the mixed strategy FPL-UE on finding the best partner from a set of opponents within the repeated opportunistic encounters between networks.

FPL-UE makes use of Geometric Resampling (GR) (Algorithm 3) in order to compute the estimated reward for the chosen action at every trial (Algorithm 4, Steps 10-11). The application of GR in our setting is shown in Algorithm 3. Basically, GR measures the reoccurrence where simulated a, denoted as ã, may appear. Thus, K_val_k represents the reciprocal of the probability of action k (p_k^{-1}), i.e. K_val provides a 1-in-M scale for probabilities, where M is a finite value that bounds the number of samples. For example, the reciprocal of 0.01 is 100, so an event with probability 0.01 has a 1 in 100 chance of happening.

Algorithm 3: Algorithm GR

Input : $M \in \mathbb{Z}^+$, $a_{i,j}(tr)$;
Output: $K\_val_k \in \mathbb{Z}^+$
1 **for** $i \leftarrow 1$ **to** M **do**
2 Repeat steps 3 ∼ 9 in Algorithm FPL-UE once to sample ã;
3 **if** $i < M$ **and** $\tilde{a} = a_{i,j}(tr)$ **then**
4 $K\_val_k = i$;
5 **else**
6 $K\_val_k = M$;
7 **end**
8 **if** $K\_val_k > 0$ **then**
9 break;
10 **end**
11 **end**

4.2.3. EXP3

Unlike FPL-UE, EXP3 employs the value of the probabilities for each action more explicitly [56]. The partner selection strategy using EXP3 is described in Algorithm 5. At each trial, EXP3 chooses a partner $a_{i,j}(tr)$ according to the distribution p (Step 4) learned from the iterations. EXP3 as FPL and ε-Greedy is a mixed strategy that introduces uniform randomisation into the action selection process. Once the action has been determined, the received reward is used to update the weight value of the chosen action (Steps 5-7), which affects proportionally to the probability of each action

Algorithm 4: Algorithm FPL-UE for each agent i, j

Input : $\lambda \in [0,1]$, $\eta \in \mathbb{R}^+$, $M \in \mathbb{Z}^+$, opponents $1, ..., K$;
Output: Negotiation partner $a_{i,j}(tr)$
1 Initialisation: $\hat{r}_k = 0$ for $k = 1, 2, ..., K$;
2 **for** $tr \leftarrow 1$ **to** Tr **do**
3 Set flag $\in \{0, 1\}$ such that $flag = 0$ with prob. λ;
4 **if** $flag$ **then**
5 $a_{i,j}(tr) = \arg\max_{k \in \{1,...,K\}} (\hat{r}_k + z_k)$; where $z_k \sim exp(\eta)$ independently for $k = 1, 2, ..., K$;
6 **else**
7 $a_{i,j}(tr) \sim U\{1, K\}$;
8 **end**
9 Receive reward $r_{a_{i,j}}(tr)$ as $E_{i,j}^{alloc}(t)$ for all t in negotiation against selected partner $a_{i,j}(tr)$;
10 Run GR($M, a_{i,j}(tr)$) to estimate p_k^{-1} as K_val_k;
11 $\hat{r}_k = \hat{r}_k + K\_val_k \cdot r_{a_{i,j}}(tr)$ where $k = a_{i,j}(tr)$;
12 **end**

in the next trial (Step 3) (i.e. the higher the current estimate is, the higher the probability an agent chooses that action). Thus, at each trial, EXP3 updates the value of the distribution p, and it defines the action with higher probability and vice versa.

Algorithm 5: Algorithm EXP3 for each agent i, j

Input : $\gamma \in [0,1]$, opponents $1, ..., K$;
Output: Negotiation partner $a_{i,j}(tr)$
1 Initialisation: $w_k = 1$ for $k = 1, 2, ..., K$;
2 **for** $tr \leftarrow 1$ **to** Tr **do**
3 Set $p_k = (1-\gamma) \cdot \frac{w_k}{\sum_{k=1}^{K} w_k} + \frac{\gamma}{K}$ for $k = 1, 2, ..., K$;
4 Draw $a_{i,j}(tr)$ randomly according to the probabilities $p_1, p_2, ..., p_K$;
5 Receive reward $r_{a_{i,j}}(tr)$ as $E_{i,j}^{alloc}(t)$ for all t in negotiation against selected partner $a_{i,j}(tr)$;
6 $\hat{r}_k = \frac{r_{a_{i,j}}(tr)}{p_k}$ where $k = a_{i,j}(tr)$;
7 $w_k = w_k \cdot \exp\left(\frac{\gamma \cdot \hat{r}_k}{K}\right)$ where $k = a_{i,j}(tr)$;
8 **end**

Although EXP3 is classified under the category of MAB algorithms with partial information, that is, only the reward of the selected action can be observed, the update of its weight affects proportionally the weights of each respective arm. According to this, EXP3 selects at each trial the best-estimated action and provides an updated probability as the

learning process continues, i.e. it guarantees an agent can efficiently adapt to different environmental situations. The exploration rate, as in FPL-UE and ϵ-Greedy is given parametrically and affects the efficiency of the algorithm as well (γ in EXP3). It is important then, to analyse first the definition of this parameter before any comparison is executed.

4.2.4. EXP3.S

EXP3.S belongs to the family of exponential weight methods. The difference with EXP3 is the boundary on its expected regret. Instead of bounding the regret with respect to the single best action, EXP3.S proves a bound for any sequence of actions. Information as the time horizon Tr and the value of "hardness" are used to determine the optimal value of the parameters α and γ in EXP3.S [56].

Algorithm 6: Algorithm EXP3.S for each agent i, j

Input : $\gamma \in [0,1]$, $\alpha = 1/Tr$, opponents $1, ..., K$;
Output: Negotiation partner $a_{i,j}(tr)$

1 Initialisation: $w_k = 1$ for $k = 1, 2, ..., K$;
2 **for** $tr \leftarrow 1$ **to** Tr **do**
3 Set $p_k = (1-\gamma) \cdot \frac{w_k}{\sum_{k=1}^{K} w_k} + \frac{\gamma}{K}$ for $k = 1, 2, ..., K$;
4 Draw $a_{i,j}(tr)$ randomly according to the probabilities $p_1, p_2, ..., p_K$;
5 Receive reward $r_{a_{i,j}}(tr)$ as $E_{i,j}^{alloc}(t)$ for all t in negotiation against selected partner $a_{i,j}(tr)$;
6 $\hat{r}_k = \frac{r_{a_{i,j}}(tr)}{p_k}$ where $k = a_{i,j}(tr)$;
7 $w_k = w_k \cdot \exp\left(\frac{\gamma \cdot \hat{r}_k}{K}\right) + \frac{e \cdot \alpha}{K} \cdot \sum_{k=1}^{K} w_k$ for $k = 1, 2, ..., K$;
8 **end**

All the notations are reported in Appendix A.

5. Experimental validation

This section describes the goal of the experiments, and the methodology followed to empirically evaluate the MAB algorithms.

5.1. Goal of the experiments

The goals of the experiments described in this paper are:

- Apply MAB learning to our setting of partner selection between multiple sensor networks for an efficient energy allocation in the long term.
- Compare four state-of-the-art strategies described above for the adversarial MAB problem presented here, using as a baseline the best-fixed strategy in hindsight, the optimal policy and the uniform random selection of a partner in each negotiation encounter.
- Evaluate through extensive simulations the performance and validate the theoretical properties of the online learning policies in a practical case study as the partner selection problem during opportunistic encounters between wireless sensor nodes under different circumstances.

5.2. Experiment scenarios

We assume four authorities that deploy their networks in the same geographic area, in such a way that there may be between three to four distinct agents within overlapping radio coverage, i.e. for each agent, there is a pool of K parties formed by 2 or 3 opponent agents from which an agent can choose one partner to initiate a bilateral negotiation. As already mentioned, in the context of partner selection for opportunistic energy negotiation, we may view agents in the pool as arms. An agent must decide between these 2 or 3 agents which arm is expected to provide the best payoff. This setup is suitable for our experiments; however, the pool of arms can be formed with any number of nodes, greater than two (depending on the memory limitations) to evaluate the MAB algorithms. The following scenarios define the changes that characterise the dynamic and heterogeneous domain of WSNs. In this paper, we focus on environmental changes such as varying energy availability, which influences the different patterns of the agent's negotiation behaviour, and also instances of network topology variation.

Our scenarios consist of general WSN applications that periodically report measurements to the sink. These networks are typically deployed for long-term operation, and their design constraints are application-dependent, also based on the monitored environment. This implies that if there is a pool of agents from which to select a negotiation's opponent, each arm will have unique characteristics that will determine its reward. This reward will depend on the negotiation outcome, which is directly affected by the negotiation strategy used by each party and their mutual zone of agreement. A mutual agreement relies on the energy availability of the arms and their ability to meet the current aspirational demand of the other agent. In addition, we are aware of topology changes during the networks' operation time. We thus focus on the participant's differences and their dynamics. Regarding the dynamic nature of these networks, besides taking into account the varying status (due to node failures, time-delays, active/sleep modes) that define their network topology, we consider changes in their attitude towards negotiation (Conceder/Boulware tactics) and environmental conditions that modify the energy availability of the agents involved. Given this, we examine the following possible situations where the MAB model is applicable:

Cooperative scenario. All the agents are Conceder negotiators. In proposed approaches to enable cooperation in multi-domain sensor networks, sensor nodes are assumed to be spontaneously cooperative. The utility function in these studies is characterised by the effective gain of minimising the nodes' energy consumption. The battery-powered networks represented in these works find that the equilibrium state with the highest payoff (where the lifetime of the sen-

sors is the highest) consists of cooperative strategies. Similarly, this first scenario assumes the Conceder strategy for the generation of offers, as a cooperative effort. Thus, the cooperative behaviour of an agent is represented here as concessions quickly performed at the beginning of the negotiation.

Multiple behaviours. The opponents adapt their negotiation behaviour according to their energy availability, which is determined by the weather conditions. If the agent requires more energy than the amount it can provide to its opponent, it adopts a tough behaviour, otherwise, it employs a Conceder strategy. The behaviour is modelled using the tactic's function (Equation (15)). In multi-authority WSNs, a resource-constrained node may be reluctant to forward packets received from other network domain, or to do any other task on behalf of an external network to save its own resources. In other words, when an agent is aware of its power level, it adapts its strategy to avoid being exploited by selfish decisions.

Dynamic topology. The networks change their topology. In this paper, we seek for an efficient learning method that finds a trade-off between exploring and exploiting the available options of opponents by jointly considering the dynamically changing environment and varying network topology. Topology changes are frequent in a sensor network and can be attributed either to node mobility, failure or node state. The networks that we are studying are not mobile networks, however, dynamic topology is considered in terms of node failures, the new addition of nodes, as well as its different states such as activeness and sleepiness.

In each scenario, the energy availability of an agent is determined by its energy harvested. Each agent in the pool of arms may have an associated reward, which corresponds to the outcome of the negotiation. This reward, as we mentioned previously, does not follow a fixed distribution, instead, it is chosen by the adversarial according to the negotiation. The negotiation outcome is defined by the mutual space of agreements and the agents' behaviour. To show the dynamism of the domain we divide the simulation time period into intervals called epochs, and each epoch lasts a number of trials Tr. Each trial involves an opportunistic negotiation interaction between two agents.

The characteristics of each party are constant along with a fixed number of trials, or epoch. At the end of an epoch, the features amongst arms change to set a different optimal partner (which is unknown by the agent during the selection). For us, a preferred opponent is one that has more energy availability and a Conceder behaviour. The feature of energy availability changes per epoch in all situations and the behaviour varies in the situations of multiple behaviours and dynamic topology. All runs in all scenarios involve 5000 trials. Four cases are considered: long epochs or static environment ($Tr = 5000$ iterations), moderately dynamic epochs ($Tr = 1000$), dynamic epochs ($Tr = 500$) and extremely dynamic epochs ($Tr = 200$). The length of epochs is obviously not known by the agents. The goal of an agent is to maximise its total reward over these trials, by finding the partner with the highest expected payoff. This determination is accomplished by observing the reward to know the efficiency of the chosen opponent, and thus, learn which opponents are the most efficient ones.

5.3. Design of experiments

Given the description of the three scenarios above, we now formulate the conditions used to alter the environment. Furthermore, we define energy availability and the behaviour of the agents in every situation. Then the topology changes are also described. The performances of the algorithms for the listed scenarios are compared and discussed in the following subsection.

We evaluate our approach experimentally using the representative scenarios described in Subsection 5.2, weather data downloaded from Weather Underground and nodes modelled to be Memsic MICAz motes. The weather information used to compute the energy generation in every experiment corresponds to the area of Southampton, UK over the year 2017 [47, 48]. The selector agent uses a different type of energy harvesting source (e.g. solar) from the agents in the pool (e.g. wind) but the values of solar irradiance and wind speed correspond to the same day, same time.

In all three scenarios, the energy harvested by the agents in the pool is modified in order to simulate environmental changes [1] that affect the performance of the energy source and generate different energy availability for each agent in the pool. These conditions determine a setting where one opponent is the best choice in every possible negotiation. Any setting with different conditions also shows the same broad patterns in the result of the simulations. The information about the characteristics of the opponents (how quickly they change per epoch and how they differ) is unknown to the agents. If there are three agents in the pool of arms, we simulate three different environmental conditions:

- Condition 1. First opponent is the best option.
  - First opponent: $E^{hrv}_{i,j}$ is not affected.
  - Second opponent: $E^{hrv}_{i,j}$ is reduced to a 40%.
  - Third opponent: $E^{hrv}_{i,j}$ is reduced to a 10%.
- Condition 2. Third opponent is the best option.
  - First opponent: $E^{hrv}_{i,j}$ is reduced to a 10%.
  - Second opponent: $E^{hrv}_{i,j}$ is reduced to a 40%.
  - Third opponent: $E^{hrv}_{i,j}$ is not affected.
- Condition 3. Second opponent is the best option.
  - First opponent: $E^{hrv}_{i,j}$ is reduced to a 40%.
  - Second opponent: $E^{hrv}_{i,j}$ is not affected.
  - Third opponent: $E^{hrv}_{i,j}$ is reduced to a 10%.

We simulate agents with two or three opponents. The agents that have two options in the set of arms will have only

[1] Energy harvested can be affected by multiple causes as obstruction of power source, weather conditions, solar panel and wind turbine efficiency.

two situations where the first or second opponent is the best option, respectively.

The **cooperative scenario** simulate the three environmental conditions described above. The strategic behaviour of each agent in the set of opponents is not affected no matter how low is its energy availability. All the agents concede more rapidly at the beginning of the negotiation with a 0.05 concession shape value.

For the **multiple behaviours scenario**, the tactic of an agent is modelled using three concession shape values (β): 0.05 for a Conceder agent, while 1.4 and 1.9 model a Boulware strategy. In this case, we follow the environmental conditions described previously that alter the energy availability but we also include varying tactics: the best option has a Conceder behaviour 0.05, while the rest of the set will have 1.4 and 1.9 respectively.

The third scenario of **dynamic topology** exhibits the conditions of the multiple behaviours scenario plus the assumptions made on the networks' topology. Every 20 trials, this simulation assigns a probability of 0.4 to allow the absence of any opponent chosen uniformly random. This represents the dynamic behaviour of the network topology, where the absence can be seen as an agent's rejection of being part of an OEN, an agent's failure, or an agent in the sleep state. As a result, every 20 opportunistic encounters, any node from the pool of opponents may be unavailable.

The duty cycle of each agent is set uniformly random between 1% to 5%, which defines its load using the power consumption model from equation (1). Finally, in order to capture the dynamic nature of the environment, we simulate the four cases described in Subsection 5.2: static characteristics over time (1 epoch), moderately dynamic changes (5 epochs), dynamic (10 epochs) and extremely dynamic case (25 epochs). When the environment changes its epoch, it uniformly randomly chooses one of the three conditions specified above. If there are two agents in the pool, then only two conditions are swapped.

The scenario setup includes a deployment area of 500 m $\times$ 500 m and randomly located sensor nodes, where each node has up to 100 m communication range. For demonstration purposes, we show the results for one of the networks involved, where 5 agents need to select a partner among a set of opponents. All simulation results correspond to the arithmetic mean of 10 simulation runs and 5000 trials each, with differences in the agents' load, number of opponents, and environmental conditions. We use as a baseline the best-fixed strategy, the optimal strategy, and random selection to measure the performance of the MAB algorithms.

5.4. Comparison of the MAB algorithms

The algorithms used in our study condition their performance on the election of an exploration rate. In order to correctly tune this parameter in EXP3, the bound on the reward (upper bound) of the best single action over all trials needs to be considered. Similarly, the optimal value of the exploration parameter for EXP3.S necessarily depends on the scenario. In particular, the time horizon Tr of performing opportunistic encounters must be known, and the same for FPL-UE. These parameters are chosen for optimising the regret bound of the policies. However, to implement the partner selection algorithms truly independent from such knowledge (e.g., one might not have Tr in advance), we are allowed to tune the exploration rate arbitrarily. Therefore, in this work, we select low exploration rates. By choosing these values, the algorithms can already achieve good results. Specifically, we set $\varepsilon = 0.1$ in ε-Greedy, $\lambda = 0.014$ in FPL-UE and $\gamma = 0.3$ in EXP3 and EXP3.S, and show by simulations how the performance of the four policies is compared to that of the optimal and best-fixed policy. Given this selection of the exploration factor, we now detail the numerical results obtained from each scenario.

5.4.1. Cooperative scenario

Fig. 5 illustrates how the agents perform using every policy described in this work in static, moderately dynamic, dynamic, and extremely dynamic environment when the topology is fixed and agents behave in a cooperative way. Fig. 5a shows ε-Greedy as the best algorithm to select a negotiation partner. During 5000 opportunistic encounters in constant competitiveness between the opponents, the most appropriate policy is a simple greedy approach. It enforces only 10% of randomness in its strategy to explore sub-optimal options, but also to consider environmental changes.

The second best choice in this configuration is FPL-UE policy. FPL-UE selects the best partner as the one that generates the maximum estimated reward over time, which in this case, is fixed for the entire duration of the particular experiment. On the other hand, EXP3 and EXP3.S use the exploration factor to maintain a list of weights for each of the opponents. These weights support the mixed strategies on deciding which action to take next. Both algorithms may benefit from their methodology if the environment changes over time, since they will use the weights to adapt to such conditions. As seen from the figure, EXP3 and EXP3.S achieve similar performance.

The selected algorithms learn to play actions that enhance the overall performance of the agents and need to be applied in the specific domain to know which strategy is the best against the corresponding problem. For instance, in the context of adversarial online learning in defender-attacker encounters, FPL-UE has proved to achieve efficient results against the best-fixed strategy on hindsight [54]. Our hypothesis, in fact, was that the negotiating agents using FPL-UE policy can tackle better the adaptive behaviour of the opponents. Although FPL-UE has certainly improved the performance of the agents, in general, the EXP3 algorithms have shown to efficiently deal with the partner selection problem in more dynamic settings (see Fig. 5b-d).

The non-learning approach or random selection of the agents consistently shows poor performance. The random selection is the worst of the benchmarks but it is implicitly suggested in the existing literature on cooperative packet forwarding. In this case, the random selection strategy in the experiments achieves up to 63% of the energy that can be

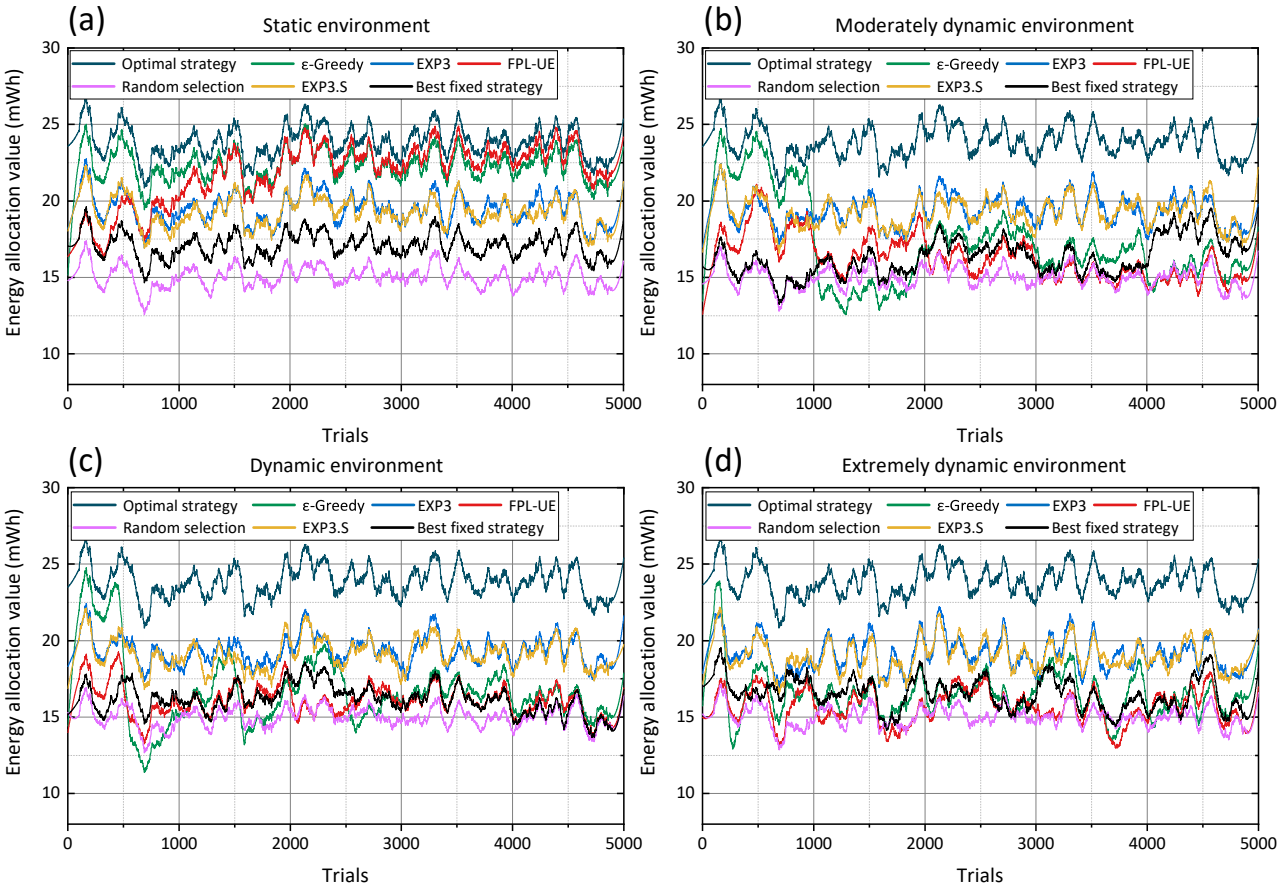

Figure 5: Cooperative scenario. Energy allocation in a 5-agent network with static network topology and Conceder agents ($\beta = 0.05$), in static, moderately dynamic, dynamic, and extremely dynamic environments.

allocated when the selection of a partner is performed intelligently in every interaction (optimal strategy). From the scenario covered in Fig. 5a, ϵ-Greedy is efficient with an average efficiency of 93%, followed by FPL-UE with 91%, EXP3 with 82% and EXP3.S with 81%, respectively. As a result, a MAB policy as ϵ-Greedy can improve up to 30% the energy allocated by an agent relative to the random selection of a partner in the static environment.

The performance of the action-selection strategies compared to that of the optimal strategy is degraded when changes appear in the environment. Their efficiency is affected even more when environmental transitions take place more frequently (see Fig. 5b-d). Temporal changes in the reward distribution structure are an intrinsic characteristic of our domain. These changes at every decision epoch vary the expectation of the rewards and motivate the agents to dismiss information gathered about the opponents, which in turn encourages exploration. However, the less time the agents have to adapt to these variations, the less they are able to characterise the reward distributions. In contrast, the theoretical properties of the online learning policies with respect to low regret bounds against the best-fixed strategy on hindsight are consistent. In fact, all the approaches have efficient theoretical performance guarantees. FPL-UE shows the worst performance compared to other approaches, but even in the most dynamic scenario, it achieves 94.77% of the energy that can be allocated with the best-fixed strategy.

In comparison with the optimal strategy, EXP3 and EXP3.S achieve the best results. In particular, EXP3 is almost consistent with 82%, 81%, 81% and 80% of efficiency among respective environments: static, moderately dynamic, dynamic and extremely dynamic. In the same manner, EXP3.S achieves 81%, 81%, 80% and 80%, respectively. The sensitivity to disturbance is more notorious in FPL-UE, where efficiency decreases up to 25% when changes occur. Similarly, ϵ-Greedy on average loses its ability to opportunistically select a negotiation partner up to 24%. Thus, the EXP3 algorithms are better in all three dynamic conditions. From these results, an agent can increase its energy allocation up to 17% using EXP3 or EXP3.S as partner selection strategy in the most dynamic environment studied in this scenario where all agents make faster concessions.

In conclusion, in this first scenario in a static environment, ϵ-Greedy policy is enough when temporal changes can be avoided over time in a partner selection problem. EXP3 and EXP3.S, however, outperform the other two policies for a broad range of temporal uncertainties in the environment.

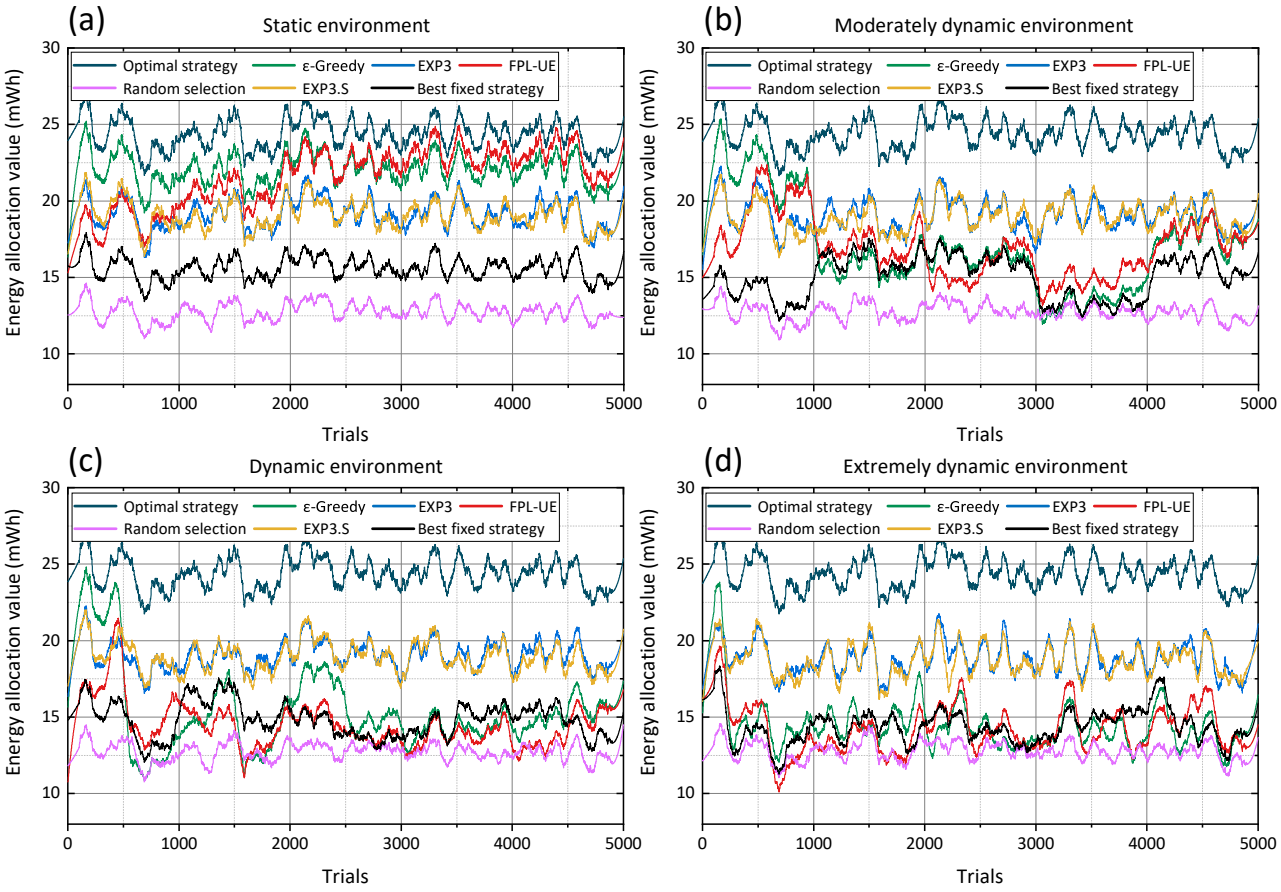

Figure 6: Multiple behaviours scenario. Energy allocation in a 5-agent network with static network topology and agents with β between 0.05, 1.4 and 1.9 in static, moderately dynamic, dynamic, and extremely dynamic environments.

5.4.2. Multiple behaviours scenario

The second scenario where agents are simulated with multiple negotiation behaviours is evaluated in Fig. 6. The negotiation behaviours, in this case, determine the target energy allocation value an agent desires in each round of the negotiation encounter. Similarly to Fig. 5, the results are shown under four degrees of environmental dynamism with respect to the energy availability that directly affects the negotiation behaviour of the agents: in static, moderately dynamic, dynamic, and extremely dynamic environments. As it can be observed in the figure, the energy allocation in average has changed in comparison to the energy allocation achieved when all the agents are Conceder. Following the results from the random selection strategy, it is observable a slight reduction. Specifically, the results obtained when all agents behave "cooperatively" report 11% more than the amount of energy allocated in this scenario (52%). In any case, even if the agents offer concessions rapidly at the beginning of the encounters, the selection of the most appropriate partner by intelligently choosing the opponent, makes a difference in our model.

Now, we can see from Fig. 6a that the performance of every approach is decreased, compared to that of the case of cooperative networks. This is, however, due to the fact that the average energy allocation amount achieved by the agent's optimal strategy has increased in the network. The main reason behind this difference is that there are fewer concessions among the agents and the desired utility levels are higher. In the cases where there are agents with a Conceder behaviour against an opponent with a Boulware behaviour, if the first one has enough energy to power itself and share, or requires a minimum amount of cooperation, the agent playing a Boulware tactic gets a better agreement. This meets the following statement: when Boulwares make deals, they receive a higher individual utility [14]. The second reason for this variation in the policies' performance is the dynamism introduced by the multiple negotiation's behaviours. In this scenario, the set of opponents offer different amounts of energy values and the diversity of potential agreements is increased between the agents. The environment is then more dynamic from an agent's perspective since the agent's behaviours change according to the amount of energy they harvest. Consequently, the variability of the opponent's negotiation tactics directly impacts the learning curve of the agents.

The adaptation to these variations is best approached by the EXP3 algorithms (EXP3 and EXP3.S). Specifically, in the static environment, the EXP3 techniques achieve up to 79% on average, of the total energy that can be allocated with

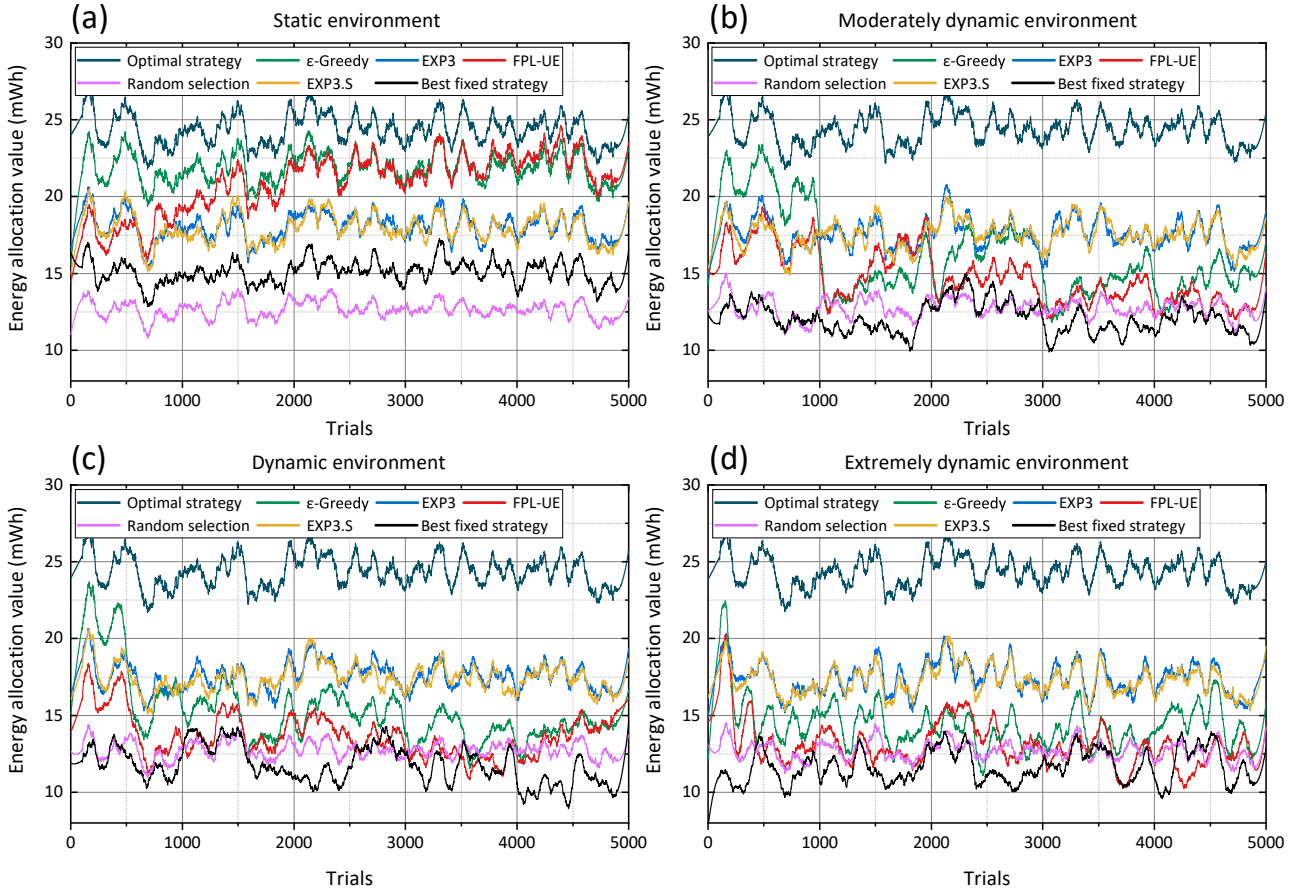

Figure 7: Dynamic topology scenario. Energy allocation in a 5-agent network with dynamic network topology and β between 0.05, 1.4 and 1.9 in static, moderately dynamic, dynamic, and extremely dynamic environments.

the optimal strategy. In comparison with the first scenario (Fig. 5a), this is only 3% less of its original capacity. Such level remains stable, as seen in Fig. 6b-d, where the efficiency of EXP3 is of 78%, 78%, and 77% for the moderately dynamic, dynamic and extremely dynamic environment, and the exact same values for EXP3.S in every environment, respectively. The other two policies reduce their performance as more dynamicity is considered in the agent's behaviours. ε-Greedy reduces its performance, on average, up to 9% in the most dynamic environment with respect to the conceders scenario (Fig. 5d), while FPL-UE policy reports a decrement up to 8% of the amount obtained in the cooperative scenario (Fig. 5d). This indicates that the EXP3 algorithms are less sensitive to the negotiation strategy changes than the rest of the policies. Most important, the EXP3 estimation method is not affected by the introduction of negotiation in the system. Overall, the learning approaches achieve better results compared to the random selection of the negotiation partner over time. Correspondingly, an agent can improve up to 25% its energy allocation when it uses the EXP3 policy against the random selection strategy of a partner in the most dynamic environment.

5.4.3. Dynamic topologies scenario

The results for the last scenario are depicted in Fig. 7. The changes in the networks' topology are taken into account while the environmental changes on weather conditions are also studied. In this regard, the energy availability affects the negotiation strategy of the agents. Thus, the agents have to deal with the challenges of environmental changes and the varying operational status of the opponents. The figure shows how the performance of the policies is again decreased by the introduction of the agent's movements (because of failure, rejection to be part of OEN, activity commute between active/sleep status). For instance, the energy amount allocated by the agent using EXP3 is reduced up to 9% in comparison to the amount allocated in the first scenario and decreased up to 6% of the average energy amount allocated when there are no topology changes but multiple negotiation behaviours. The same occurs with the rest of the policies on a different level.

For the FPL-UE policy and ε-Greedy, the efficiency is reduced up to 12% and 10% from the results obtained in the cooperative scenario (Fig. 5). In reference to the multiple behaviours scenario without topology changes (Fig. 6), the allocation is discounted up to 9% using FPL-UE and 4% using ε-Greedy. These results show that the EXP3 algo-

rithms represent again the best solution as the environmental changes become more frequent (see Fig. 7b-d, respectively). In fact, an agent deciding a partner using the EXP3 algorithms achieves 71% - 72% efficiency while the use of ε-Greedy supplies 59% efficiency and FPL-UE learning approach obtains 54% in the most challenging case studied in our work, i.e. in extremely dynamic environments with environmental and topology changes. The efficiency of 72% achieved by EXP3 increases by up to 20% the energy allocation achieved using a random selection strategy.

Despite the fact that the agents' performance using EXP3 is affected by the topology changes, this policy achieves the best results with respect to the optimal strategy, the best-fixed strategy and the random selection. EXP3 approach is consistent through the performance evaluation in each scenario and its reward estimation method proved to handle more realistic domains of complex and dynamic environments. This is supported by the results depicted in the variety of scenarios studied here. Moreover, EXP3 is not sensitive to the negotiation strategies incorporated in the decision process. Thus, the adaptive learning feature provided by EXP3 is the most suitable solution for the problem of partner selection in our domain. Furthermore, the EXP3 policy can be applied in a broader range of negotiation agents interactions where computationally-lightweight solutions are required. The results of our research are quite useful for designing agents in open environments that need to cope with the uncertainty of the adversarial setting and network conditions. In this case, the MAB learning model presented in our work allows an agent to select the most prospective partner from a set of opponents and reach efficient energy allocation agreements in the long term.

In the next section, we discuss the cost of the establishment of OEN when agents need to discover the set of opponents in their vicinity.

6. Establishing the OEN

Before the agents face the challenge of selecting a negotiation partner, they need to discovery the negotiation agents in the neighbourhood, i.e. the agents that want to cooperate and establish an opportunistic energy negotiation. This section evaluates the cost associated with the overhead of the discovery protocol in the network's performance. Results show the average energy consumption of 50 simulation runs with different network topologies.

OEN adopts a publish-subscribe approach in which the agents conserve energy by sending a limited amount of messages. Three types of messages are exchanged between agents: OEN_ADV, OEN_REQUEST and OEN_ACCEPT. Initially, the agents are deployed with a cross-domain link-layer protocol as OI-MAC [41] and a standard routing protocol. OEN is implemented between the link and network layers. In this way, it takes advantage of their functionalities: the agents can communicate directly with co-located agents using the capabilities of the link layer protocol, while still are able to inform the network layer about the cooperative

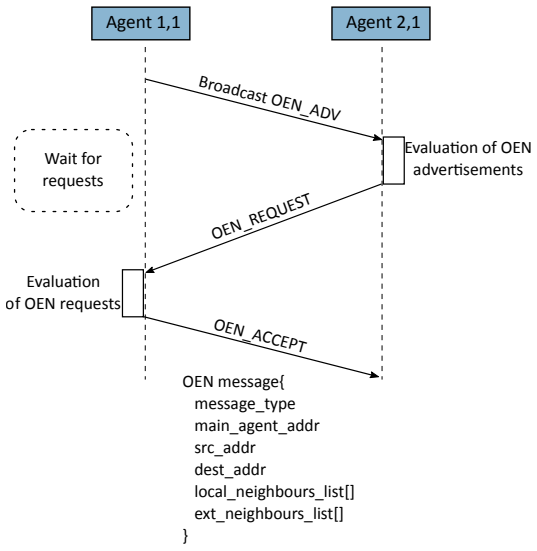

Figure 8: Sequence diagram of OEN establishment.

agreements reached with the counterparts. In the WSNs cooperation literature, the networks increase their performance by cooperative packet forwarding. We use this specific type of cooperation to guide the decision process of being part of an opportunistic energy negotiation.

Once the agents are deployed, the first step in OEN for a negotiator is to broadcast through its immediate neighbours on all available radio frequencies, the desire to start a negotiation by sending an OEN_ADV message. From that moment, the agent becomes the main agent: the agent that will choose a negotiation partner from a set of opponents. OEN_ADV includes the list of agents in its range (from the local and external network domain) and a query to find the neighbours of the neighbouring agent contacted. Figure 8 illustrates the discovery protocol in a sequence diagram and the OEN header format.

At this point, there are two possible situations per neighbouring agent reached. The main agent (call this agent $1, 1$) with another agent (call this agent $2, 1$) may have no interaction. The information provided about the nodes in range by agent $1, 1$ may not be ideal for agent $2, 1$ and it can simply ignore the main agent's request. Thus, agent $1, 1$, after waiting for a certain interval of time, drops the communication with agent $2, 1$ and stays in the initial state while the number of nodes discarded is different from the total number of its neighbours. On the other hand, agent $2, 1$ may accept the main agent's proposal. In this situation, agent $2, 1$ sends an OEN_REQUEST using the radio frequency that is associated with the main agent $1, 1$ to ask for participation in OEN. In this message, agent $2, 1$ informs agent $1, 1$ about the agents that are in its range.

Again, there are two possible scenarios. First, agent $1, 1$ may ignore agent $2, 1$'s request. This may happen for two reasons: agent $1, 1$ is already part of an OEN with another set of agents or agent $1, 1$ is now unreachable. Agent $2, 1$ then waits for a grace period before discarding agent

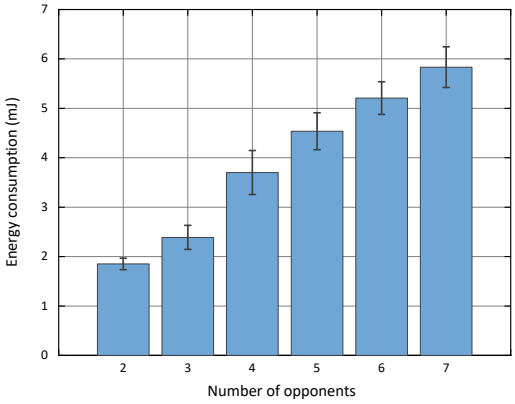

Figure 9: Average energy spent at the end of 6 seconds plotted against the number of opponents reached.

1, 1's proposal. The second possible scenario includes a response. Agent 1, 1 may accept the agent's request and send an OEN_ACCEPT message to add agent 2, 1 to the pool of opponents. This leads agent 1, 1 to a selection state if the number of agents in the set of opponents is bigger than one, if not, the agent moves to a final state, the state of negotiation. In the state of negotiation, both agents can directly establish a bilateral negotiation. Conversely, in the selection state, agent 1, 1 employs an action-selection policy (EXP3 is the best strategy according to our results) in order to select one negotiation partner from the pool of agents and move to the final state of negotiation.

The discovery protocol was tested using OMNeT++ [57]. The effects of the OEN discovery protocol are evaluated on energy consumption. The simulation setup includes 5, 10, 15, 20 and 25 overlapped sensor nodes randomly deployed in an area of 100 m × 100 m over 50 simulation runs for each density. PHY and MAC layers are defined by the IEEE 802.15.4 standard, while the rest of the parameters used in these simulations are summarised in Table 1.

Table 1
Simulation parameters for nodes' power usage in OMNeT++.

Parameter	Definition
Standard	IEEE 802.15.4
Simulation time	6 s
Tx current	17.4 mA
Idle listen current	0.02 mA
Rx current	18.8 mA
Rx-Tx current	0.02 mA
Voltage	3 V

Fig. 9 shows the average energy cost of transmission of an agent against 2 to 7 opponents during the simulation period. As can be seen, the OEN discovery protocol consistently consumes more energy when the pool of opponents is increased.

The discovery protocol, however, has an insignificant impact on energy consumption (< 0.01 J), and is a result of the continuous reception required for negotiation agents discovery. Once the agent broadcasts an advertisement message, it listens to recept the request messages of the neighbouring networks. During this process, the agents share details to decide whether they should associate with a possibility of cooperation, depending on the contribution each could give to the opponent network (by exchanging only context information). This step aligns the goals of the individual agents to find compatibility, thus ensuring that the networks can self-organise into communities deciding how to cooperate, through a negotiation mechanism. We propose that our negotiation-based cooperation approach can facilitate the interaction and collaborative management in a wide range of applications and can lead to an efficient coexistence of multiple co-located networks. For instance, the relatively minor increase in energy consumption is likely to be outweighed by the 59% increase in energy allocation that an agent may achieve when it reaches an agreement with a strategically selected partner from a different network domain.

7. Conclusions and future work

In this work, we proposed a novel partner selection model based on multi-armed bandit learning, that allows each agent of a network to adaptively optimise its operation by the selection of a partner that maximises its energy allocation. We have also applied a negotiation framework to model the interaction between agents based on a time-dependent tactic and orthogonal strategy in a resource-constrained domain with incomplete information. The negotiation techniques employed in our work allows co-located devices to decide a cooperation effort while handling the uncertainty of the environment. Note that in such experiments, an agent's utility increases up to 59% with OEN.

In this paper, we extend the state-of-the-art in cooperation between networks by reporting on a negotiation-based mechanism to address the preferences conflict of highly heterogeneous agents. The 5 steps of the methodology that guide the cooperation are adopted to accomplish a specific goal: opportunistic energy negotiation, called OEN. Since networks can have different or multiple optimisation goals, the proposed phases can be custom-tailored towards a specific objective.

With the aim to optimise a network's power management using the suggested approach, the first step for an agent is to identify its own efficiency. In the domain of OEN, it corresponds to the energy allocation scheme that a node can employ to power its load. Thus, the optimal energy allocation algorithm described in Subsection 3.2 is proposed. This energy allocation scheme is evaluated during every simulation presented in this work. Such algorithm enables self-organised agents that can anticipate insufficient energy allocation schemes and the opportunity to start an OEN.

For the partner selection problem, we proposed a multi-armed bandit based approach that reduces the complexity of address reasoning to negotiate. For instance, instead of focusing on the selection of a partner based on the benefits

associated with it (as approached in our model), the negotiation might involve strategies to model the negotiation behaviour of the opponents. In particular, some strategies include regression techniques to estimate the concessions of the adversaries and predict possible agreements. Since we are interested in resource-constrained domains, we concentrate on low complexity solutions that don't require complex learning mechanisms. The predicting techniques use a sufficient number of the opponent's offers to apply the learning approach and start the estimation of the counterpart's information (such as its deadline or reservation values) in order to obtain better deals. In fact, the complexity of the utility space increases with the interdependent issues and the number of time slots involved in the energy cooperation domain.

Against this background, we described state-of-the-art MAB algorithms applied in the negotiation context, ε-greedy, EXP3, EXP3.S and FPL-UE, as the bandit strategies for efficiently deal with the uncertainty regarding the preference of the opponents and the dynamism of the environment. Our results show that even in a cooperative scenario, where agents offer concessions rapidly at the beginning of the encounters, the agents improve their benefit by choosing a partner strategically instead of select it randomly. In average up to 17% more energy can be allocated in an extremely dynamic environment using the EXP3 policy. The problem becomes even more challenging as we target setups where agents employ tougher negotiation strategies and the presence of the agents is unstable. In any case, the bandit strategies achieve improved energy allocation agreements by adjusting to dynamic environments against the random selection of a negotiation partner. The results showed up to 25% more energy allocation over the random selection strategy in the multiple behaviours scenario and up to 19% increment in the dynamic topologies scenario. In this direction, the EXP3 policy produces better results at a large number of unexpected events as the environment becomes more dynamic.

The establishment of OEN proved to have a minimum impact on energy cost. Using a discrete event simulator as OMNeT++, the discovery protocol to reach the negotiation agents in a 1-hop neighbourhood is implemented as a publish-subscribe protocol. The simulation included up to 7 opponents subscribed to the OEN process, where a negligible impact in the agent's performance lower than 0.01 J was found. Thus, the obtained results demonstrate that a node can engage in OEN with a minimum cost even in the emergence of seven co-located and distinct nodes.

As future work, the discovery of agents with a desire for cooperation can be reduced by choosing reliable agents found in previous interactions. In particular, we would like to extend this work to reduce the system overhead of finding compatible agents to negotiate by using previous experiences and acting in consequence. In addition to this, the introduction of multiple incentives (such as low delay, higher coverage, better QoS guarantees) is expected to improve the efficiency of the negotiation agreements. A tradeoff negotiation approach can outperform a concession approach in terms of utility but may incur a highest communication cost and computational power. The analysis of the combination of these models is also left as future work.

Appendix A. Notations

Symbol	Measure	Description
K	–	set of arms or actions
$u_{i,j}$	Wh	utility function of agent i, j
Tr	–	total number of opportunistic negotiation encounters
tr	–	the tr-th encounter, tr = 1,2,...,Tr
$a(tr)$	–	action of an agent at trial tr
$r_{i,j}$	Wh	reward function of agent i, j
R_{Tr}	Wh	weak regret over Tr
ε	–	exploration factor in ε-Greedy action selection strategy
$\hat{r}_k$	Wh	estimated reward of action k
$pulls_k$	–	number of times action k has been executed in ε-Greedy
$rewards_k$	Wh	cumulative reward of action k in ε-Greedy
λ	–	exploration factor in FPL-UE
η	–	mean parameter for exponential distribution in FPL-UE
M	–	maximum number of samples in FPL-UE
K_val_k	–	reciprocal of probability of action k in FPL-UE
z_k	–	exponential random number in FPL-UE
γ	–	exploration factor in EXP3 and EXP3.S
w_k	–	weight value of action k in EXP3 and EXP3.S
p_k	–	probability of action k in EXP3 and EXP3.S
α	–	learning parameter in EXP3.S

Acknowledgments

This research was supported in part by the Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), Escuela Superior Politécnica del Litoral (ESPOL) under Ph.D. studies 2016, and the UK Engineering and Physical Sciences Research Council (EPSRC) under Platform Grant EP/P010164/1.

Data supporting the results presented in this paper is available at DOI: https://doi.org/10.5258/SOTON/D1659.

References

[1] A. P. Ortega, G. V. Merrett, S. D. Ramchurn, Automated negotiation for opportunistic energy trading between neighbouring wireless sensor networks, in: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), IEEE, 2018, pp. 1–6.

[2] S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, arXiv preprint arXiv:1204.5721 (2012).

[3] A. K. et al., Power management in energy harvesting sensor networks, ACM Transactions on Embedded Computing Systems vol. 6 (2007) 32.

[4] D. K. Noh, L. Wang, Y. Yang, H. K. Le, T. Abdelzaher, Minimum variance energy allocation for a solar-powered sensor system, in: International Conference on Distributed Computing in Sensor Systems, Springer, 2009, pp. 44–57.

[5] S. Chen, P. Sinha, N. B. Shroff, C. Joo, A simple asymptotically optimal joint energy allocation and routing scheme in rechargeable sensor networks, IEEE/ACM Transactions on Networking 22 (2013) 1325–1336.

[6] M. H. Anisi, G. Abdul-Salaam, M. Y. I. Idris, A. W. A. Wahab, I. Ahmedy, Energy harvesting and battery power based routing in wireless sensor networks, Wireless Networks 23 (2017) 249–266.

[7] B. Zhang, R. Simon, H. Aydin, Maximum utility rate allocation for energy harvesting wireless sensor networks, in: Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems, ACM, 2011, pp. 7–16.

[8] Y. Zhang, S. He, J. Chen, Data gathering optimization by dynamic sensing and routing in rechargeable sensor networks, IEEE/ACM Transactions on Networking 24 (2015) 1632–1646.

[9] A. M. et al., A negotiation protocol for multiple interdependent issues negotiation over energy exchange, in: Proc. of the AI for an Intelligent Planet, 2011, p. 1.

[10] L. Buttyán, T. Holczer, P. Schaffer, Spontaneous cooperation in multi-domain sensor networks, in: European Workshop on Security in Ad-hoc and Sensor Networks, Springer, 2005, pp. 42–53.

[11] M. Félegyházi, J.-P. Hubaux, L. Buttyán, Cooperative packet forwarding in multi-domain sensor networks, in: Third IEEE International Conference on Pervasive Computing and Communications Workshops, IEEE, 2005, pp. 345–349.

[12] F. Garcin, M. H. Manshaei, J.-P. Hubaux, Cooperation in underwater sensor networks, in: Game Theory for Networks, 2009. GameNets' 09. International Conference on, IEEE, 2009, pp. 540–548.

[13] N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra, M. Wooldridge, Automated negotiation: prospects, methods and challenges, International Journal of Group Decision and Negotiation 10 (2001) 199–215.

[14] P. Faratin, C. Sierra, N. R. Jennings, Negotiation decision functions for autonomous agents, Robotics and Autonomous Systems 24 (1998) 159–182.

[15] X. Zheng, P. Martin, K. Brohman, L. Da Xu, Cloud service negotiation in internet of things environment: A mixed approach, IEEE Transactions on Industrial Informatics 10 (2014) 1506–1515.

[16] B. An, V. Lesser, D. Irwin, M. Zink, Automated negotiation with decommitment for dynamic resource allocation in cloud computing, in: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 981–988.

[17] R. Zheng, N. Chakraborty, T. Dai, K. Sycara, M. Lewis, Automated bilateral multiple-issue negotiation with no information about opponent, in: 2013 46th Hawaii International Conference on System Sciences, IEEE, 2013, pp. 520–527.

[18] R. Zheng, N. Chakraborty, T. Dai, K. Sycara, Multiagent negotiation on multiple issues with incomplete information, in: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 1279–1280.

[19] A. Diamah, M. Wagner, M. van den Briel, A comparative study on vector similarity methods for offer generation in multi-attribute negotiation, in: Australasian Joint Conference on Artificial Intelligence, Springer, 2015, pp. 149–156.

[20] L. Niu, F. Ren, M. Zhang, Feasible negotiation procedures for multiple interdependent negotiations, in: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 641–649.

[21] P. Faratin, C. Sierra, N. R. Jennings, Using similarity criteria to make issue trade-offs in automated negotiations, artificial Intelligence 142 (2002) 205–237.

[22] D. Somefun, E. H. Gerding, S. Bohte, J. A. La Poutré, Automated negotiation and bundling of information goods, in: International Workshop on Agent-Mediated Electronic Commerce, Springer, 2003, pp. 1–17.

[23] G. Lai, K. Sycara, A generic framework for automated multi-attribute negotiation, Group Decision and Negotiation 18 (2009) 169.

[24] M. Wu, M. de Weerdt, H. La Poutré, Efficient methods for multi-agent multi-issue negotiation: Allocating resources, in: International Conference on Principles and Practice of Multi-Agent Systems, Springer, 2009, pp. 97–112.

[25] M. Rovcanin, E. De Poorter, I. Moerman, P. Demeester, A reinforcement learning based solution for cognitive network cooperation between co-located, heterogeneous wireless sensor networks, Ad Hoc Networks 17 (2014) 98–113.

[26] M. A. Alsheikh, S. Lin, D. Niyato, H.-P. Tan, Machine learning in wireless sensor networks: Algorithms, strategies, and applications, IEEE Communications Surveys & Tutorials 16 (2014) 1996–2018.

[27] P. Wang, T. Wang, Adaptive routing for sensor networks using reinforcement learning, in: The Sixth IEEE International Conference on Computer and Information Technology (CIT'06), IEEE, 2006, pp. 219–219.

[28] R. Arroyo-Valles, R. Alaiz-Rodriguez, A. Guerrero-Curieses, J. Cid-Sueiro, Q-probabilistic routing in wireless sensor networks, in: 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, IEEE, 2007, pp. 1–6.

[29] L. Tran-Thanh, A. Rogers, N. R. Jennings, Long-term information collection with energy harvesting wireless sensors: a multi-armed bandit based approach, Autonomous Agents and Multi-Agent Systems 25 (2012) 352–394.

[30] J. Zhu, Y. Song, D. Jiang, H. Song, Multi-armed bandit channel access scheme with cognitive radio technology in wireless sensor networks for the internet of things, IEEE access 4 (2016) 4609–4617.

[31] L. Li, J. Ren, Q. Zhu, On the application of lora lpwan technology in sailing monitoring system, in: 2017 13th Annual Conference on Wireless On-demand Network Systems and Services (WONS), IEEE, 2017, pp. 77–80.

[32] K. Mekki, E. Bajic, F. Chaxel, F. Meyer, A comparative study of lpwan technologies for large-scale iot deployment, ICT express 5 (2019) 1–7.

[33] R. Kerkouche, R. Alami, R. Féraud, N. Varsier, P. Maillé, Node-based optimization of lora transmissions with multi-armed bandit algorithms, in: 2018 25th International Conference on Telecommunications (ICT), IEEE, 2018, pp. 521–526.

[34] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, J. Palicot, Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings, in: International Conference on Cognitive Radio Oriented Wireless Networks, Springer, 2017, pp. 173–185.
[35] H. Dakdouk, E. Tarazona, R. Alami, R. Féraud, G. Z. Papadopoulos, P. Maillé, Reinforcement learning techniques for optimized channel hopping in ieee 802.15. 4-tsch networks, in: Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2018, pp. 99–107.
[36] S. Maghsudi, S. Stańczak, Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework, IEEE Transactions on Vehicular Technology 64 (2014) 4565–4578.
[37] S. Munroe, M. Luck, Motivation-based selection of negotiation opponents, in: International Workshop on Engineering Societies in the Agents World, Springer, 2004, pp. 119–138.
[38] S. S. Fatima, M. Wooldridge, N. R. Jennings, The influence of information on negotiation equilibrium, in: International Workshop on Agent-Mediated Electronic Commerce, Springer, 2002, pp. 180–193.
[39] S. Radu, E. Kalisz, A. M. Florea, A model of automated negotiation based on agents profiles, Scalable Computing: Practice and Experience 14 (2013) 47–56.
[40] J. Brzostowski, R. Kowalczyk, On possibilistic case-based reasoning for selecting partners for multi-attribute agent negotiation, in: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, ACM, 2005, pp. 273–279.
[41] T. J. et al., Opportunistic direct interconnection between co-located wireless sensor networks, in: Int. Conf. on Computer Communications and Networks, 2013, pp. 1–5.
[42] G. Weiss, A modern approach to distributed artificial intelligence, IEEE transactions on systems man & cybernetics-part c applications & reviews 22 (1999).
[43] K. Singhanat, N. R. Harris, G. V. Merrett, Experimental validation of opportunistic direct interconnection between different wireless sensor networks, in: 2016 IEEE Sensors Applications Symposium (SAS), IEEE, 2016, pp. 1–6.
[44] X. Jiang, J. Polastre, D. Culler, Perpetual environmentally powered sensor networks, in: Proceedings of the 4th international symposium on Information processing in sensor networks, IEEE Press, 2005, p. 65.
[45] J. F. Nash Jr, The bargaining problem, Econometrica: Journal of the Econometric Society (1950) 155–162.
[46] A. Rubinstein, Perfect equilibrium in a bargaining model, Econometrica: Journal of the Econometric Society (1982) 97–109.
[47] Southampton Weather, https://www.wunderground.com/, 1995. URL: https://www.wunderground.com/, accessed on April 01, 2019.
[48] PVGIS, http://re.jrc.ec.europa.eu/, 2001. URL: http://re.jrc.ec.europa.eu/, accessed on April 01, 2019.
[49] A. A. Babayo, M. H. Anisi, I. Ali, A review on energy management schemes in energy harvesting wireless sensor networks, Renewable and Sustainable Energy Reviews 76 (2017) 1176–1184.
[50] S. R. et al., Power sources for wireless sensor networks, in: European workshop on wireless sensor networks, 2004, pp. 1–17.
[51] H. Robbins, Some aspects of the sequential design of experiments, Bulletin of the American Mathematical Society 58 (1952) 527–535.
[52] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction (2011).
[53] A. Kalai, S. Vempala, Efficient algorithms for online decision problems, Journal of Computer and System Sciences 71 (2005) 291–307.
[54] H. Xu, L. Tran-Thanh, N. R. Jennings, Playing repeated security games with no prior knowledge, in: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 104–112.
[55] G. Neu, G. Bartók, An efficient algorithm for learning with semi-bandit feedback, in: International Conference on Algorithmic Learning Theory, Springer, 2013, pp. 234–248.
[56] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The non-stochastic multiarmed bandit problem, SIAM journal on computing 32 (2002) 48–77.
[57] A. Varga, Omnet++, in: Modeling and tools for network simulation, Springer, 2010, pp. 35–59.