IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 13, 2021, accepted June 4, 2021, date of publication June 7, 2021, date of current version June 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3087410

User-Centric Radio Access Technology Selection: A Survey of Game Theory Models and Multi-Agent Learning Algorithms

GIUSEPPE CASO[1], (Member, IEEE), ÖZGÜ ALAY[1,2], (Member, IEEE), GUIDO CARLO FERRANTE[3], (Senior Member, IEEE), LUCA DE NARDIS[4], (Member, IEEE), MARIA-GABRIELLA DI BENEDETTO[4], (Fellow, IEEE), AND ANNA BRUNSTROM[5], (Member, IEEE)

[1]Department of Mobile Systems and Analytics, Simula Metropolitan Center for Digital Engineering, 0167 Oslo, Norway
[2]Department of Informatics, University of Oslo, 0373 Oslo, Norway
[3]Ericsson Research, 164 40 Kista, Sweden
[4]Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, 00184 Rome, Italy
[5]Department of Computer Science, Karlstad University, 651 88 Karlstad, Sweden

Corresponding author: Giuseppe Caso (giuseppe@simula.no)

The work of Giuseppe Caso, Özgü Alay, and Anna Brunstrom were supported in part by EU Horizon 2020 Research and Innovation Program [5th Generation End-to-End Network, Experimentation, System Integration, and Showcasing (5GENESIS)] under Agreement 815178.

ABSTRACT User-centric radio access technology (RAT) selection is a key communication paradigm, given the increased number of available RATs and increased cognitive capabilities at the user end. When considered against traditional network-centric approaches, user-centric RAT selection results in reduced network-side management load, and leads to lower operational costs for RATs, as well as improved quality of service (QoS) and quality of experience (QoE) for users. The complex between-users interactions involved in RAT selection require, however, specific analyses, toward developing reliable and efficient schemes. Two theoretical frameworks are most often applied to user-centric RAT selection analysis, i.e., game theory (GT) and multi-agent learning (MAL). As a consequence, several GT models and MAL algorithms have been recently proposed to solve the problem at hand. A comprehensive discussion of such models and algorithms is, however, currently missing. Moreover, novel issues introduced by next-generation communication systems also need to be addressed. This paper proposes to fill the above gaps by providing a unified reference for both ongoing research and future research directions in the field. In particular, the review addresses the most common GT and MAL models and algorithms, and scenario settings adopted in user-centric RAT selection in terms of utility function and network topology. Regarding GT, the review focuses on non-cooperative models, because of their widespread use in RAT selection; as for MAL, a large number of algorithms are described, ranging from game-theoretic to reinforcement learning (RL) schemes, and also including most recent approaches, such as deep RL (DRL) and multi-armed bandit (MAB). Models and algorithms are analyzed by comparatively reviewing relevant literature. Finally, open challenges are discussed, in light of ongoing research and standardization activities.

INDEX TERMS Radio access technology selection, game theory, multi-agent learning, reinforcement learning.

I. INTRODUCTION

Nowadays, communication devices are most often equipped with multiple radio access technologies (RATs), and this feature is set to increase in the future. Users are thus able to connect to several RATs, also referred to as heterogeneous networks (HetNets), including wireless wide, metropolitan, local, personal, and body area networks (WWANs, WMANs, WLANs, WPANs, and WBANs). The development of HetNets aims at providing multiple and heterogeneous services, including end-to-end communication and data exchange through the Internet, *anytime*, *anywhere*, and at reasonable

The associate editor coordinating the review of this manuscript and approving it for publication was Tiankui Zhang.

levels of quality of service (QoS) and experience (QoE) to *anyone* [1], [2].

Meeting QoS and QoE requirements, that depend on the requested service, involves actuating the always best connected (ABC) paradigm, including the action of selecting the RAT(s) to be connected to [3], [4]. RAT selection is therefore widely investigated by research and standardization communities, and typically refers to HetNets selection, user association, offloading, and horizontal/vertical handover (or handoff) mechanisms [5]–[10] (cf. Section III). *Network selection* is also often used to generically indicate RAT selection [4]–[6].

RAT selection schemes differentiate depending upon *how*, *where*, and *why* ABC procedures are executed [10]. In a classic approach, dedicated controllers perform selection at the network side (*how* and *where*), resulting in centralized network-centric solutions, leading most often to a system optimal configuration (*why*). In other approaches, the selection is performed either cooperatively between RAT access nodes and user devices, or entirely at the user side (*how* and *where*). These solutions, referred to as hybrid and distributed user-centric schemes, address the scalability issues of centralized schemes, while putting more emphasis on user satisfaction (*why*). A review of centralized vs. hybrid vs. distributed RAT selection can be found in [9]–[11].

Centralized approaches often result in higher system performance; there is however an increasing interest toward user-centric schemes, due to two main reasons [4], [9]–[16]:

Densification: The ongoing exponential growth of heterogeneous access nodes and user devices challenges the possibility to solve the selection problem by one or a few network controllers. This is due to the high computational complexity of finding a system-level optimal solution and the huge increase of signaling messages from (to) the controllers.

Cognition: The increased cognitive and computational capabilities of user devices enable autonomous rational decisions based on *context-awareness* at the user end, i.e., by observing and adapting to the surrounding *context*. This is in line with recent trends toward the decentralization of several network functionalities and decisions, as currently proposed in edge and fog networking [17].

A. THEORETICAL FRAMEWORKS FOR RAT SELECTION

The theoretical analysis of RAT selection has been addressed by adopting several approaches. For example, RAT selection has been often modeled as an optimization problem, where the main goal is to maximize a system utility function across network entities (e.g., users and RATs), under a set of constraints that depend on how the problem is formulated (e.g., maximize system throughput under possible resource constraints of RATs). Assuming that users can connect to a single RAT at a time, the problem is combinatorial and NP-hard. In order to find a near-optimal solution, the connection constraint is usually relaxed [9], and the resulting convex problem can be solved, e.g., via Lagrangian dual analysis [18]–[21], divide-and-conquer [22], and learning approaches [23], [24]. In most cases, the analysis is only partially focused on user-centric RAT selection schemes, since part of the optimization is usually solved at network side (e.g., see [18], where the problem is decomposed in two sub-problems solved by running dedicated algorithms at user and network sides). This is particularly true when RAT selection is solved jointly with resource allocation (also, *user scheduling*), that addresses the problem of how RATs allocate their resources (e.g., time and/or frequency) to connected users [22], [23].

This paper focuses on user-centric RAT selection, as defined earlier in this section. The analysis involves the interaction among end users, that may or may not cooperate in the actuation of RAT selection strategies, in order to maximize (minimize) either their own or the overall system *utility* (*cost*)[1] function, which include QoS and/or QoE parameters. The utility of users is affected by their own selection strategies and may also be affected by the surrounding context, e.g., radio conditions. A context retrieval procedure is thus required to let each user *learn over time* the relationship between context, strategies, and utility. The goal is to converge to a set of strategies that optimize utility, while driving the system into a *stable* configuration.

Two complementing theories are most widely adopted in user-centric RAT selection analysis:

Game Theory (GT): GT is mainly used as a framework for *modeling* RAT selection scenarios. GT models the interaction among rational decision makers, referred to as *players* or *agents*, having common or conflicting utility interests and adopting either cooperative or non-cooperative strategies. The set of GT models, in terms of both players' behavior and observed/shared information, provides a framework under which distributed optimization problems can be analyzed. In these scenarios, players may have different goals and partial control over the system [25], [26].

Multi-Agent Learning (MAL): MAL is mainly used as a framework for *solving* RAT selection games. Rooted in single-agent learning (SAL), MAL provides algorithmic solutions for the process (by each agent (*learner*)) of discovering and adapting to the surrounding context (including other agents). MAL algorithms define policies adopted by learners to interact with one another and with the context, and aim at the optimization of utility [27].

The mapping between players and learners reveals the relationship between GT and MAL. The key point of a joint GT-MAL analysis is that, if a GT solution exists, that is, there exists an *equilibrium* as defined in non-cooperative GT, the GT solution can be achieved by applying a MAL algorithm. Hence, the challenge of analyzing so-called *learning games* is to demonstrate the existence of solutions by

[1] From now on, the analysis is mostly carried out in terms of utility, switching to cost whenever appropriate. Due to the different terms traditionally used in game theory and multi-agent learning, *payoff*, *reward*, and *return* are also used to identify utility or related functions, while *regret* may be used to refer to cost. Changes in definitions are made explicit whenever needed.

GT models, and adopt a reliable, practical, and scalable MAL algorithm to reach them [28]–[30].

The applicability of MAL algorithms depends on the game model in terms of both players' behavior and observed/shared information [31]. For non-cooperative games, when a player knows in advance the features of the surrounding context, such as the utility and how it is affected by strategies of others, its iterative process focuses on *learning the equilibria* [32], [33], i.e., finding a stable system configuration given a pre-known context. In this case, algorithms derived from game-theoretic analyses, e.g., best response dynamics (BRD) and fictitious play (FP), can be adopted [27, Chapter 5] [31, Chapter 5]. The assumption of a known context while learning equilibria allows to adopt the hypothesis of *full rationality* of players, i.e., players exclusively act to optimize their utility [34], [35].

When the initial knowledge of the context is limited, so-called multi-agent *reinforcement* learning (MARL) solutions can be adopted, derived from single-agent reinforcement learning (SARL) [27, Chapter 5] [31, Chapter 6] [36], [37]. SARL is a branch of machine learning (ML) addressing agent adaptation to its surrounding context (often referred to as *environment*), that is unknown at interaction kick-off [38]–[40]. SARL algorithms solve so-called Markov decision process (MDP) and multi-armed bandit (MAB) scenarios, by balancing *exploration* vs. *exploitation* strategies over time, i.e., alternating learning vs. utility optimization strategies.

The idea behind MARL is that players learn the equilibria *while* learning the context, e.g., by exploring all possible strategies. In scenarios of extremely limited observable information, players adopting MARL may even be unaware of being part of a game [36]. The need for learning context and equilibria leads to the hypothesis of *bounded rationality* of players [35].

B. CONTRIBUTION

GT has been largely used to analyze the behaviour of wireless communications [41]. Leveraging its application to single-RAT scenarios for distributed channel selection [42], GT has also been used for distributed multi-RAT selection. More recently, MAL algorithms have also been used to show the capability of users to solve RAT selection in a distributed manner, under the hypothesis of initial limited knowledge of their surrounding context.

Therefore, a large amount of literature focuses on analyzing, modeling, and proposing methods for user-centric RAT selection in a joint GT-MAL framework. A comprehensive analysis of the proposed, but at times conflicting, GT models and MAL algorithms is, however, missing. Moreover, novel issues are raised by next-generation communication systems and scenarios, which require further investigation toward efficient and practical strategies.

This paper aims at providing a unified reference on user-centric RAT selection from both GT and MAL perspectives. In particular, the paper surveys *users vs. users games for user-centric RAT selection*,[2] where end users are the main entities that interact, most often, indirectly with one another during the selection process.

The contribution of this paper can be summarized as follows:

- We provide a primer on RAT selection, by analyzing key concepts and standardized mechanisms in use in current wireless communication systems;
- We analyze and describe the most common non-cooperative GT settings proposed for modeling user-centric RAT selection. We also discuss practical scenario aspects, i.e., adopted utility functions and network topologies;
- We analyze and describe the most common MAL algorithms proposed for solving user-centric RAT selection, including game-theoretic, RL, DRL, and other schemes. The literature is reviewed and analyzed, with particular focus on user-centric RAT selection;
- We highlight the need for a multi-faceted performance evaluation of RAT selection, by providing a taxonomy of the main performance indicators to be considered in order to provide exhaustive analyses;
- We discuss open challenges and possible future work, in light of the ongoing evolution of GT, MAL, and communication systems.

The proposed GT-MAL perspective allows for a critical literature review, and ultimately suggests further refinements toward addressing future challenges.

C. STRUCTURE

The paper is organized as follows. Section II provides the background of the present work. First, it gives an overview of applications of GT and MAL to wireless communications; then, it introduces existing surveys and tutorials that discuss, to some extent, the application of GT and MAL to RAT selection, ultimately comparing such investigations with the contribution of this paper. Section III provides the foundations of RAT selection, by discussing key concepts and standardized mechanisms. Two aspects that are relevant to the analysis of RAT selection by GT-MAL, i.e., the definition of utility and the choice of a network topology, are also analyzed. Section IV summarizes game-theoretic aspects adopted for modeling RAT selection, while Section V describes three non-cooperative models often used as RAT selection games, and also provides an initial literature review. Figures 1a and 1b provide a detailed content summary of Sections IV and V. Section VI focuses on MAL, and describes learning algorithms for solving RAT selection, and further refines the literature review. Figure 2 provides a detailed content summary of Section VI. A description of the indicators used to analyze the performance of learning schemes is provided in Section VII, while Section VIII reviews open challenges and possible future work. Section IX concludes the paper.

[2] *(User-centric) RAT selection* is simply used in the following.

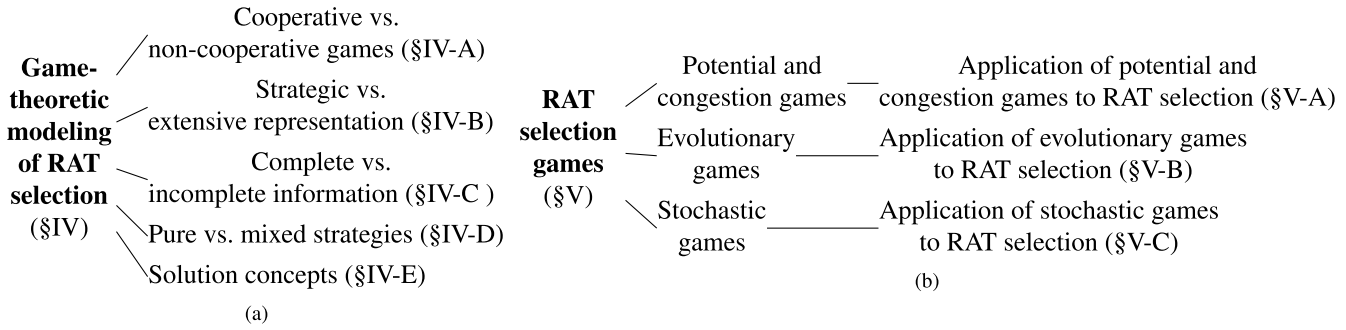

FIGURE 1. Content summary of Section IV (a) and Section V (b).

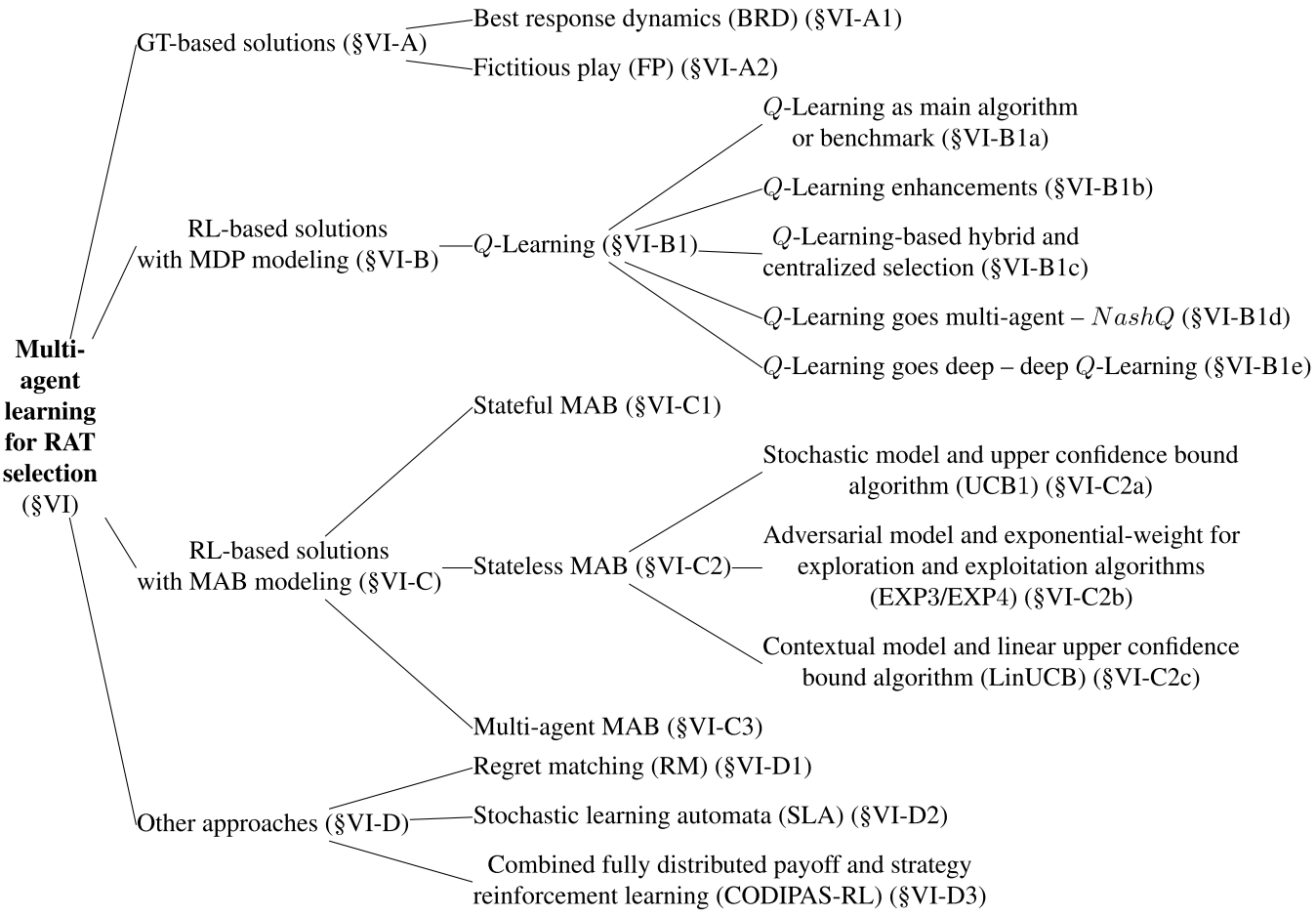

FIGURE 2. Content summary of Section VI.

II. BACKGROUND AND RELATED WORK

A. GT AND MAL APPLIED TO WIRELESS COMMUNICATIONS

In order to grasp the theoretical connections between GT and MA(R)L with no specific focus on wireless communications, the interested reader may refer to [27], [36], [37], [50]–[53], among the vast literature on this subject. Furthermore, for general assessments on the application of RL, in its deep version (DRL) [54], [55], to multi-agent systems, the reader may also refer to recent works such as [56], [57].

For detailed analyses on the use of GT and MAL in wireless communications and signal processing, it is useful to refer to notable books, such as [31], [58], [59] and [60]. Reference [61] presents a comprehensive overview of game-theoretic tools applied to wireless communications, mainly focusing on *static* games, i.e., games where the interaction is limited to a single iteration, and thus learning is not considered. A discussion on *dynamic* games and MAL, with focus on signal processing applications, is provided in [32]; BRD, FP, regret matching (RM), and RL algorithms are briefly described. The same algorithms are reviewed in [33], where they are applied to a 2-player game between two transmitter-receiver pairs, that aim at not interfering with one another, while exchanging data on a common set of frequency

TABLE 1. Key literature related to modeling and analysis of user-centric RAT selection, also with respect to the present work.

Reference (Year)	GT	MAL	Further Comments
[5] (2012)	Discussion of cooperative and non-cooperative GT models. Literature taxonomy in users vs. users, users vs. networks, and networks vs. networks	Not discussed	Discussion of utility functions, multi-operator scenarios, pricing and energy aspects.
[6] (2013)	Mentioned as a modeling option	Not discussed	Comparison of several modeling approaches. Focus on utility definition. Learning partially discussed from a single-agent perspective through MDPs.
[10] (2017)	Limited discussion	Focus on a baseline algorithm [43] [44]	Analysis of network-assisted mechanisms, noisy context retrieval, and mmWave scenarios.
[16] (2018)	Limited discussion	Focus on a selection of RL algorithms	Comparative analysis against examples of centralized and hybrid approaches. Analysis in terms of network topology and user density. Adoption of a specific utility function.
[45] (2019)	Used for modeling RAT selection in specific situations [45, Part III]	Focus on RL algorithms (with MDP and MAB models) [45, Parts I-II]	Review of a selection of literature works (including [46]–[49]). DRL and QoE-based utility models mentioned as future work.
This work (2021)	**Focus on non-cooperative GT: modeling options and game models (§IV-§V)**	**Discussion of several MAL schemes: GT-based, (D)RL-based, and further approaches (§VI)**	**Focus on users vs. users scenarios. Discussion on utility function and network topology. Review of standardized mechanisms. Discussion on performance indicators and open challenges. Comprehensive literature review and comparative analysis (§IV-§VIII).**

channels. More recently, [62] discusses the application of DRL to multi-agent scenarios in wireless communications.

B. GT AND MAL APPLIED TO RAT SELECTION

This section summarizes the state-of-the-art in terms of surveys and tutorials that focus specifically on RAT selection, in order to frame the context and further highlight the contribution of this paper.

Reference [5] presents an extensive literature review of GT-based RAT selection, covering cooperative vs. non-cooperative games and several other modeling aspects. It classifies the literature into users vs. users, networks vs. users, and networks vs. networks games. This taxonomy is reused for example in [6]. Reference [5] also discusses open challenges related to the definition of utility function, to multi-operator and multi-technology scenarios, and to pricing and energy consumption aspects. Basic concepts on GT are provided, but MAL aspects are not analyzed (covered in this work, in Section VI).

Reference [6] reviews the mathematical theories adopted to model user-centric RAT selection, including Utility and Markov decision theories, fuzzy logic, combinatorial optimization, GT, and multiple attribute decision making (MADM), a branch of the more general multiple criteria decision making (MCDM). GT is presented as one among several modeling theories, but the aspect of learning the equilibria while discovering and adapting to the context is beyond the scope of [6]. In addition, the analysis of the Markov decision theory, which includes MDPs and MABs, is carried out from a single-player perspective (this work expands to a multi-agent perspective, via the definition of stochastic games and corresponding learning schemes, cf. Sections V-C and VI).

User-centric RAT selection is also discussed in [10], where a baseline selection algorithm is introduced, building upon [43], [44]. The work analyzes aspects related to network-assisted mechanisms, millimeter-wave (mmWave) networks, and the effect of noisy measurements during context retrieval, but provides a limited discussion on GT-MAL joint analysis (covered in this work, in Sections IV-VI).

More emphasis on MAL is given in [16], that reports a comparative analysis between centralized, hybrid, and distributed algorithms. Variable network topology, user density, and a specific utility function are considered, aiming at providing a fair comparison between algorithms (also made available as a software library). On the same line, a recent review of selected works, with corresponding models and learning schemes, is provided in [45]. These same works are also discussed in this paper, and analyzed in comparison with other existing literature.

Table 1 summarizes the above references, and compares them against the contribution of the present work.

Other investigations discuss practical aspects rather than theoretical modeling and analyses. Focusing on 5th generation (5G) cellular networks, user association is surveyed in [9], in terms of metrics, topology, and control; GT is reported as a modeling choice along with combinatorial optimization and stochastic geometry. Literature overviews and in-depth discussions on practical aspects related to offloading and handover mechanisms can be found in [7], [8], [63], and [64].

III. FUNDAMENTALS OF RAT SELECTION

Before moving to the analysis of GT and MAL in the context of RAT selection, this section introduces the fundamentals

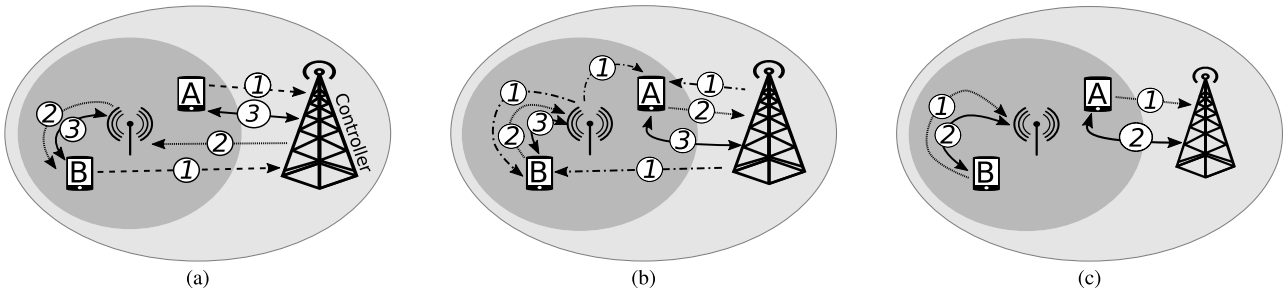

FIGURE 3. A representation of *centralized* (a), *hybrid* (b), and *distributed* (c) RAT selection schemes. All cases assume that both users A and B are initially connected to a higher priority RAT, represented as a cellular tower. The other available RAT is represented as an access point with smaller coverage. In (a), dashed lines (label 1) represent the transmission of measurement reports from users to the controller (co-located with the highest priority RAT), while small-dashed lines (label 2) represent the controller decision about the RAT users should connect to. In (b), dash-pointed lines (label 1) represent the transmission of general information (e.g., congestion levels) from RATs to users. In (b)(c), small-dashed lines (label 2 in (b), label 1 in (c)) represent the decision of users about the RAT to connect to. In (a)(b)(c), full lines (label 3 in (a)(b), label 2 in (c)) represent the establishment of a data link between users and selected RATs.

of RAT selection, by providing key concepts and examples, along with a review of standardized mechanisms. We also discuss two aspects that are relevant for GT-MAL analyses of RAT selection, i.e., the definition of a utility function and the choice of a network topology.

A. KEY CONCEPTS AND EXAMPLES OF RAT SELECTION

As anticipated in Section I, RAT selection schemes can be grouped in centralized, distributed, and hybrid, depending on which network entities perform the selection.

Centralized schemes leverage dedicated network controllers. In order to take tailored decisions, such controllers require periodic reporting from users in terms of their coverage status and experienced performance. Hence, the ongoing network densification challenges this approach and requires more scalable solutions. Distributed and hybrid mechanisms put end devices in charge of deciding the RAT(s) to connect to. This may happen in a fully independent manner (distributed), i.e., users collect their own information and perform RAT decisions independently, or via network assistance (hybrid), i.e., users also exploit indications shared from the access points of the available RATs. In the second case, these indications are usually not tailored on users and represent general context information on congestion, load, and expected performance of the RATs, as well as selection policies to be preferably applied, and that may help users to take better decisions. Both distributed and hybrid cases are thus more scalable than centralized ones, where user-specific information is continuously reported at the network side. A representation of RAT selection operations is shown in Figure 3, which provides an example of centralized (Figure 3a), hybrid (Figure 3b), and distributed (Figure 3c) selection with two users and two RATs. All cases assume that both users A and B are initially connected to a higher priority RAT, represented as a cellular tower. In Figure 3a, the highest priority RAT also acts as the controller regulating users connection across the two RATs. Users transmit the results of their context retrieval (*measurement reports*) to the controller (label 1, dashed lines), which then decides on the RAT the users should connect to (label 2, small-dashed lines); finally, user A is instructed to remain on the highest priority RAT while user B connects to the other RAT (label 3, full lines). In Figure 3b, RATs share with users general indications on their status (e.g., congestion levels) (label 1, dash-pointed lines), and users use this information and their own measurements to decide the RATs they should camp on (label 2, small-dashed lines). In Figure 3c, users only use their own measurements to decide the RATs they should camp on (label 1, small-dashed lines). In both Figures 3b and 3c, user A decides to remain on the highest priority RAT, while user B selects the other RAT (label 3 in (b), label 2 in (c), full lines).

Major standardization entities are currently defining RAT selection but also aggregation (aka multihoming) schemes that enable users to use multiple RATs in parallel. Several efforts toward enabling seamless usage of WWANs/WMANs (e.g., 4G, 5G, and WiMAX), and WLANs/WPANs (e.g., WiFi), are underway. In the following, we refer to selection and aggregation schemes as RAT *interoperability* procedures.

Regarding the 3rd generation partnership project (3GPP) cellular systems, initial standardization has focused on how and when to actuate *horizontal handover* between macrocells, that is, the transition of a user – also called user equipment (UE) – across same-tier cellular access nodes.[3] For 4G systems, including Long Term Evolution (LTE) and LTE-Advanced (LTE-A), further attention has been given to cross-tier handover between macro and small cells (contained in Release 12 (Rel-12, 2015)).

In terms of RAT aggregation, coordinated multi-point (CoMP) has been introduced in Rel-11 (2012), where multiple cells can transmit (receive) the same data toward a UE, in order to improve the communication quality in poor coverage areas. While CoMP lies across the physical and Medium Access Control (MAC) layers, dual connectivity (DC) is

[3]Macrocell is a term widely used to identify a traditional cellular access node covering medium-to-large areas. Over the years, aiming at deploying multi-tier radio access networks (RANs), macrocells have been flanked by small cells, which provide medium-to-low coverage. Depending on the coverage size, small cells are categorized in femto, pico, and micro cells.

another aggregation scheme performed in the above Packet Data Convergence Protocol (PDCP) layer. Standardized in Rel-12, DC allows UEs to exploit two not co-located LTE cells, e.g., two evolved Node Bs (eNBs). The Master eNB is part of the control plane toward the LTE Core, and coordinates with the Secondary eNB to provide additional radio resources to UEs. For 5G, initial proposals in Rel-15 (2018) have led to extending DC to support a parallel use of LTE and 5G New Radio (NR) access nodes, via several options of so-called 5G Non-Standalone (NSA) deployments [65].

All of the above cases represent centralized RAT interoperability schemes. Indeed, UEs perform and report measurements related to strength and quality of cell signals, e.g., in terms of reference signal received power (RSRP [dBm]), reference signal received quality (RSRQ [dB]), and signal-to-interference plus noise ratio (SINR [dB]). Then, cells act as controllers and coordinate on dedicated control channels, e.g., via X2 interfaces, ultimately deciding on the necessity of handover/aggregation for each UE. This is done by comparing current UE conditions with so-called handover/aggregation *trigger events*, defined in 3GPP specifications [66]–[70].

The focus is also on enabling interoperability between 3GPP and non-3GPP RATs, such as WiFi. In this context, *vertical handover* is the term commonly used to indicate the transition of a user between RATs having different priorities (e.g., from cellular to WiFi). When the handover goal is to decrease congestion on the high-priority network and better balance users on available RATs, the selection process is also referred to as *offloading* (note that, besides cellular to WiFi, vertical handover and offloading are terms that also indicate handovers between macro and small cells, as described above). In this context, 3GPP introduced a user-centric mechanism via the Access Network Discovery and Selection Function (ANDSF) in Rel-8 (2008) [71]. ANDSF is an optional element in the cellular Core, providing context information on non-3GPP systems to UEs, ultimately promoting informed selection of available WiFi networks. Considering the so far low commercial interest in ANDSF, 3GPP introduced two network-centric solutions in Rel-13 (2016), and extended in Rel-14 (2017), i.e., LTE-WLAN Aggregation (LWA) and LTE-WLAN radio-level integration with Internet Protocol (IP) security tunnel (LWIP) [72]. LWA and LWIP couple cellular and non-cellular systems at the radio level, with WiFi access points (APs) having similar functions of cellular access nodes.[4] APs and cells can be either co-located or not; in either cases, UEs report WiFi-related measurements to the cellular network, that decides whether activating the interoperability option. LWA is tailored for cellular integration with trusted WiFi networks; hence, both aggregation and selection functionalities are available via so-called *split-bearer* and (slow vs. fast) *link-switching* modes. Full offloading toward WiFi is also provided via the *switch-bearer* mode. LWIP targets integration with untrusted WiFi APs, and thus only enables slow selection and offloading.

Similar mechanisms are envisioned for 5G [73], although current standardization activities seem to be oriented toward different approaches. In particular, the above mechanisms are deployed at the radio layer, and this solution may be cumbersome for 3GPP/non-3GPP interoperability, due to the heterogeneity of resource allocation and modulation schemes across RATs. Hence, 3GPP introduced the Access Traffic Steering, Switching, and Splitting (ATSSS) architecture in Rel-16 (2020) [74]. ATSSS exploits the capability of the 5G Core of explicitly dealing with non-3GPP traffic, via the Non-3GPP Inter-Working Function (N3IWF) defined in Rel-15; a Multi-Access Protocol Data Unit is defined and exchanged over multiple RATs, in both selection (steering and switching) and aggregation (splitting) functionalities, which are deployed below or above the IP layer. In the second case, ATSSS exploits interoperability mechanisms at the transport layer, where RATs usually exploit common protocols. These solutions have been largely investigated and, as a result, several transport protocols have a *multipath* (MP)[5] extension. For example, Transmission Control Protocol (TCP) has a MP extension referred to as MPTCP, standardized by the Internet Engineering Task Force (IETF) [75]. Moreover, Concurrent Multipath Transfer (CMT) [76] and Multipath QUIC (MPQUIC) (currently in two main versions under discussion) [77] extend Stream Control Transmission Protocol (SCTP) and QUIC, respectively.

The use of MPTCP has been proposed in ATSSS [74], [78], with further discussions also pointing at the adoption of MPQUIC [79]. In all cases, MP functionalities are deployed between the UE and the 5G Core, that essentially decides and informs users on the policy to adopt for RAT interoperability. Such a use of MP transport protocols is denoted as *Core-centric*, and delineates ATSSS as a hybrid RAT selection solution, since the 5G Core decides on the selection policy to adopt (e.g., among Active-Standby, Priority-based, Smallest Delay, and Load-balancing [74]), leaving however the actual RAT decision to the connection end points (users and servers). A more general use of MP transport protocols, referred to as *Above-the-Core*, does not require the cellular Core to be involved in the policy decision. As a matter of fact, the Above-the-Core integration enables transparent interoperability, not specifically regulated by the cellular network [72], [78], with MP functionalities deployed at the communication end points, ultimately representing an example of distributed RAT selection. As further discussed in the next sections, there exist several investigations related to distributed RAT selection, also in the context of MP transport protocols (cf. Section VIII). However, "*select WiFi when detected*" is still the predominant RAT selection policy, although it is often sub-optimal in terms of QoS and QoE. An enhancement of this simple strategy is the WiFi Assist

[4]LWA/LWIP-based WiFi aggregation is regulated by the cellular network. WiFi traffic is exited and reabsorbed from/into the LTE system via LWA Adaptation Protocol (LWAAP) and LWIP Extension Protocol (LWIPEP), so that it can be handled by the LTE Core.

[5]In this context, the available RATs are often referred to as *paths*.

solution, introduced by Apple in 2016, that automatically switches back to the cellular network when WiFi is not able to meet service requirements.

The above overview of existing RAT selection schemes highlights that there is room for novel user-centric RAT selection schemes.

B. UTILITY

1) GENERAL CONCEPTS

The utility function characterizes RAT selection approaches, particularly in user-centric scenarios. Several investigations focused on proposing utility functions for RAT selection, as also reported in [5], [6]. Most of the definitions originate in the so-called Utility Theory, where utility is a mathematical formulation of the level of satisfaction of a decision-maker with respect to a particular service, when one or more *attributes* representing that service are taken into account [80]. Formal representations proposed for example in [81]–[85] emphasize that the utility function in RAT selection can be either user-specific or user-agnostic, depending on the need for differentiating users in terms of requirements and expected performance. Moreover, a utility function can include one or multiple network attributes, such as available bandwidth, connection price, and energy consumption, to mention a few.

When utility uses a single network attribute, literature reports of wide use of monotonic functions (e.g., sigmoidal and linear) [6]. When multiple attributes are considered, MADM and MCDM techniques are more often used to combine weighted and normalized attributes under a global definition, leading to methods, such as, simple additive weighting (SAW), multiplicative exponential weighting (MEW), gray relational analysis (GRA), and technique for order preference by similarity to ideal solution (TOPSIS) [82], [83], [86]–[91]. The analytical hierarchy process (AHP) approach is usually adopted for deriving weights associated to network attributes (e.g., see [87], [92], where AHP is used in GRA). In both single and multi-attribute cases, the assumption is that end devices observe network attributes and use them to evaluate a global utility.

GT highlights a further aspect that is somehow blurred in the above discussion: the dynamics of most of the attributes characterizing the RATs, e.g., the available bandwidth, are affected by the strategies adopted by users during the selection process, and so is utility. Therefore, the hypothesis of devices being capable of observing several attributes and draw the relationship between strategies of other devices, attributes, and utility, may be challenging for those devices with limited observation and computation capabilities. As clarified in Section IV, this case is well modeled by so-called imperfect and incomplete information games. Here, it is most often assumed that end devices collect *sampled values* of the utility, for which the relationship with (observable or not) attributes may be unknown. For this reason, the modeling assumption of imperfect and incomplete information, most commonly adopted in RAT selection games, avoids high-level, multi-attribute definitions of the utility, and maps the latter with an observable QoS parameter. As detailed in Section III-B2, the downlink (DL) throughput experienced by a connection to a RAT is usually adopted as the instantaneous utility associated to a strategy (e.g., a selected RAT) [44], [93].

Slightly more complex utility functions that assume higher degrees of observability or a-priori knowledge at the user side have also been proposed, e.g., utility functions based on a linear combination of DL throughput and other network attributes, such as pricing and billing costs [94]–[96] and energy consumption [97]. Energy aspects have also been considered in terms of either energy efficiency, usually defined as the ratio between user data rate and energy consumption (e.g, see [98], [99] for a system level perspective), or energy consumption savings (e.g., see [100], [101]).

2) ADOPTED UTILITY FUNCTIONS

When RAT selection is analyzed in a GT-MAL framework, the adopted utility is usually mapped onto a single QoS parameter, e.g., DL throughput. Higher-level, multi-attribute utility representations may challenge fully distributed solutions and require to move to hybrid approaches, since information signaled from other network entities, such as the access nodes of candidate RATs, may be needed for evaluating the utility. For example, three different cost functions are compared in [102] aiming at casting RAT selection as a congestion game (CG, see Section V-A). Being directly related to the number of users connected to each RAT and their throughput, the proposed functions require each user to retrieve this information, e.g., via network assistance. A similar model has been applied in [103], among others.

Defining the utility as the experienced DL throughput allows to avoid information exchange. However, in order to provide realistic simulations, a proper model is still required. To this aim, WLANs and WWANs throughput models, for example IEEE 802.11 (WiFi) [104] and cellular systems [105], are often used in RAT selection. In [44], the two following models for the instantaneous throughput experienced by user n connected to RAT a_n, referred to as M1 and M2, are provided:

$$\text{M1:} \quad r_{n,a_n} = \mathsf{f}_{a_n}(\Phi_{1,a_n}, \ldots, \Phi_{n,a_n}, \ldots, \Phi_{N_{a_n},a_n}) \tag{1a}$$

$$\text{M2:} \quad r_{n,a_n} = \mathsf{f}_{a_n}(N_{a_n}) \times \Phi_{n,a_n} \tag{1b}$$

$$\text{for all } n \in \mathcal{N}_{a_n},\ a_n \in \mathcal{A}_n.$$

In (1a)(1b), r_{n,a_n} is the throughput of user n connected to RAT a_n, and $\mathcal{A}_n$ is the set of RATs available to user n. $\mathcal{N}_{a_n}$ identifies the set of N_{a_n} users selecting candidate RAT a_n (including user n), and Φ_{n,a_n} represents the *physical rate*, i.e., the maximum achievable throughput by user n when it is the only one connected to RAT a_n. Such a value depends on radio conditions and adopted modulation and coding schemes [43], [44], [93]. Finally, function f_{a_n} is the same across users but may differ across RATs.

As also discussed in [44], [93], the M1 model resembles a RAT adopting throughput fair (TF) user scheduling, e.g., WiFi. M2 resembles instead proportional fair (PF) mechanisms, such as time division multiple access (TDMA) and orthogonal frequency division multiple access (OFDMA), as adopted for example in 4G and 5G cellular RATs. Hence, M1 and M2 can be specified as follows:

$$\text{TF:} \quad r_{n,a_n} = \Big(\sum_{k \in \mathcal{N}_{a_n}} \frac{1}{\Phi_{k,a_n}} \Big)^{-1} \tag{2a}$$

$$\text{PF:} \quad r_{n,a_n} = \frac{\Phi_{n,a_n}}{N_{a_n}} \tag{2b}$$

$$\text{for all } n \in \mathcal{N}_{a_n},\, a_n \in \mathcal{A}_n.$$

Besides providing the models, [44] also explores the impact of noisy measurements. The analysis is later extended in [93]. Indeed, (2a)(2b) provide instantaneous, *nominal* throughput, the values of which are in practice influenced by the dynamicity of physical and radio environments, e.g., user mobility and channel fading. A Gaussian distribution with mean equal to r_{n,a_n} and standard deviation $\sigma = e \times r_{n,a_n}$ (with e between 0 and 1) is thus adopted in order to consider this effect.

The models in (2a)(2b) are widely adopted in RAT selection learning games (see Section VI). Furthermore, slightly different versions of the PF model have also been proposed. On the one hand, *weighted* PF is suggested in [46], where users congest RATs with different weights, depending on their traffic and applications. Moving a step toward QoE, [46] proposes to translate throughput into mean opinion score (MOS) values, by adopting the mapping given in [106]. On the other hand, the PF model is also considered at channel granularity, e.g., in [107], where the assignment of frequency channels is embedded in the RAT selection game, and thus the physical rate also depends on how many frequency channels are available and assigned to each user.

A model similar to PF is adopted in [96] and reused in [95], [108], [109], among others. In this case, the user-specific physical rate Φ_{n,a_n} is replaced by the *capacity* of the candidate RAT, denoted as C_{a_n} and defined as the maximum data rate sustained by the RAT, independently of users. This leads to a user-independent utility definition, that complies with evolutionary GT (eGT) modeling and, being very general, also simplifies the analysis when RAT selection games are modeled as CGs, potential games (PGs), and stochastic games (see throughout Section V). A logarithmic function on top of such a simplified PF is applied in [94]–[96]. A model that also considers the assignment of frequency channels to users, and thus possible inter-user interference, is used instead in [23].

As anticipated in Section III-B1, throughput-based utility is sometimes complemented with other observable attributes, leading to novel, mostly linear, utility representations. Pricing and billing costs are commonly used in order to account for operators and service providers. Indeed, network operators may want to discourage new connections in case of saturated RATs, by charging new connections with a price proportional to the number of users [96], [109]. A-priori fixed prices can be used for differentiating RATs from a cost perspective [94], [95]. Energy aspects are also often considered in defining utility functions. In [101], energy consumption of candidate RATs, in terms of DL transmissions and management of UE handovers, is considered as utility and fed back to UEs so that those can account for it in selecting efficient policies. Among others, energy efficiency, defined as the ratio of data rate and power consumption, is considered in [22], where energy efficiency at both UE and link levels is considered for solving both RAT selection and user scheduling. In the context of energy harvesting networks, [110] defines wasted energy as the (harvested) energy used for unsuccessful transmissions. Models for both energy harvesting and energy consumption are given in [111].

C. NETWORK TOPOLOGY

The choice of a topology has key implications on both theoretical models and practical scenarios of RAT selection. Figure 4 shows the most commonly adopted topologies, i.e., *Corridor* (or *Chain*), *Overlapping*, and *Nest* [47], [112].

As regards GT models, different topologies lead to different sets of strategies $\mathcal{A}_n$ for users involved in the game (i.e., for all $n \in \mathcal{N}$). For example, in the Corridor topology of Figure 4a, users in Zone 1 and Zone 2 may have different strategy sets: the ones in Zone 1 can select between the left access node and the middle one, while the ones in Zone 2 can select between the middle access node and the right one. Users still participate in the same selection game since their strategies affect each other. A common assumption is to assign a same strategy set across users, that is, $\mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}_N$. This corresponds to analyzing the game in Zone 2 of the Overlapping topology of Figure 4b.

The selection of a network topology also leads to considering specific RAT selection mechanisms. As discussed in [47], which is a notable example where the proposed RAT selection game and corresponding MAL schemes are evaluated for the three topologies of Figure 4, the Corridor topology well represents a roadside deployment of cellular macrocells *or* small cells, while the Overlapping structure generalizes the Corridor, with macrocells *and* small cells randomly deployed with overlapping coverage areas. These two scenarios are commonly used to analyze cellular horizontal handover or user association across WiFi APs. The Nest topology (Figure 4c) better represents more heterogeneous scenarios in terms of available RATs, with a large area covered by higher tiers of WWAN or WMAN systems, and smaller regions where other RATs with smaller coverage are also available, e.g., small cells and WiFi APs. Hence, this topology is preferable for the analysis of vertical handover (e.g., cellular to WiFi or macrocell to small cell offloading).

IV. GAME-THEORETIC MODELING OF RAT SELECTION

This section describes the most common game-theoretic settings used for modeling RAT selection in realistic scenarios.

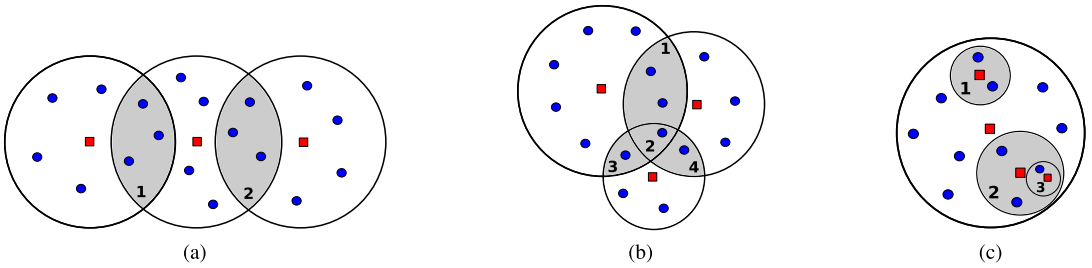

FIGURE 4. Main topologies adopted in RAT selection analysis: (a) *Corridor*, (b) *Overlapping*, and (c) *Nest*. Blue dots, red squares, and black circles identify users, candidate RATs, and RATs coverage areas, respectively. Grey areas highlight the zones where users could actually perform RAT selection, since they detect more than one RAT. Different numbers in the grey areas indicate that a different number and nature of RATs are available, and thus the users have a different set of connection strategies.

Less commonly adopted settings are also reported and briefly discussed for completeness.

A. COOPERATIVE VS. NON-COOPERATIVE GAMES

User-centric RAT selection has been modeled via both *cooperative* [113] and *non-cooperative* [114] games. Given the widespread use of the non-cooperative case, this section and the overall paper focuses on the non-cooperative case. A brief and preliminary discussion on cooperative games, with examples of application to RAT selection, is also provided for completeness.

Players *consciously* help each other in cooperative games. Cooperation can be conveyed in a game in many different ways. In *team* games, players have a common goal and coordinate with one another in order to achieve the goal; in *bargaining* games, players bargain with one another by selecting one among several possible collaboration strategies; in *coalition* games, that deal with the formation of subsets of players (coalitions) in a coordinated manner. Other examples are *matching* games, that model the problem of matching players in distinct sets depending on their information and preferences [115], [116], and *bankruptcy* games, where the focus is on finding solutions for optimal allocation of a resource across players when such a resource is not sufficient to satisfy all players' demand.

Cooperative games are usually devoted to evaluate fairness, stability, and efficiency of decision-making processes by which players are able to communicate and coordinate. The solutions that are commonly adopted in cooperative games include the core, the Shapley value, Nash bargaining, and two-sided stable matching [117, Chapters 26-31]. The application of cooperative GT to wireless communications requires communication among network entities and possibly extensive message exchange; this may invalidate the advantage of distributed solutions over centralized schemes. Even though cooperative games have been in general used to a lesser extent than non-cooperative models, for RAT selection scenarios, relevant work in the context of cooperative user association and resource allocation can be found for example in [118]–[124].

Non-cooperative games involve selfish players, whose interests and strategies may negatively affect the utility experienced by other players. In these games, players act independently, with the goal of maximizing their own utility, with no reason though to harm others. Some sort of cooperation is thus possible in a non-cooperative setting, e.g., via tailored learning schemes, by designing utility functions that are inversely proportional to a metric of selfishness, and also by including in the game the knowledge of common information to help users coordinate with one another (in RAT selection games, this may be achieved by means of *network assistance* mechanisms, as detailed in Section IV-C). As discussed in the following sections, non-cooperative games can either be *strategic* or *extensive* (Section IV-B), and consider *complete* vs. *incomplete* information (Section IV-C). Players may adopt either *pure* or *mixed* strategies (Section IV-D). Finally, these games are usually analyzed in terms of stable configurations, known as equilibria (Section IV-E).

B. STRATEGIC VS. EXTENSIVE REPRESENTATION

Non-cooperative RAT selection has been most often represented as a *strategic* game, also known as *normal* or *simultaneous* game. A strategic game $\mathcal{G}$ is characterized by the following features:

1) The set $\mathcal{N} = \{1, \ldots, N\}$ of players, that corresponds to the set of users in a RAT selection game;
2) The set $\mathcal{A}_n$ of available strategies to player n (for all $n \in \mathcal{N}$). In the simplest case, this set corresponds to the set of RATs available to user n. According to a GT common notation, a generic strategy for user n within its set $\mathcal{A}_n$ is denoted by a_n. Different subscripts stand for different users.[6] These strategies are also referred to as *pure* in GT; as discussed in Section IV-D, *mixed* strategies extend pure strategies in a probabilistic manner. Both pure and mixed strategies are used in RAT selection games;
3) The utility, that is associated with the combinations of players' strategies. It measures the *gain* obtained by each user when adopting a selection strategy, that is when selecting one RAT, as a function of the selection made by the other players. The utility for player n is

[6] Whenever needed, an explicit superscript will be added so to differentiate strategies, e.g., $a_n^i \neq a_n^j$, with $a_n^i, a_n^j \in \mathcal{A}_n$.

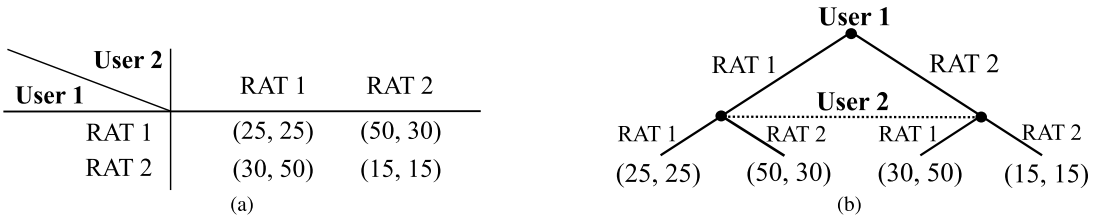

FIGURE 5. Example of a game of imperfect information: the 2-user 2-RAT selection scenario. Strategic and tree forms are given in (a) and (b), respectively. In (a), rows identify the strategies of User 1, while columns report the strategies of User 2. Both users can select either RAT 1 or RAT 2. In (b), it is assumed that User 1 moves first, before User 2; imperfect information, that is the uncertainty of User 2 on RAT selection by User 1, is represented by a dashed line connecting the vertices representing User 2. The two vertices connected by the dashed line form the information set of User 2. In both (a) and (b), for each strategy profile, that is, for all $a = (a_1, a_2) \in \mathcal{A}_{\text{tot}}$, the utilities of the two users are reported in the form (u_1, u_2).

thus defined on $\mathcal{A}_{\text{tot}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ and takes values on $\mathbb{R}$, that is, $u_n : \mathcal{A}_{\text{tot}} \to \mathbb{R}$ (for all $n \in \mathcal{N}$).

Altogether, a strategic game is commonly indicated as $\mathcal{G} := \{\mathcal{N}, \{\mathcal{A}_n\}_{n\in\mathcal{N}}, \{u_n\}_{n\in\mathcal{N}}\}$. Given a specific strategy combination $\boldsymbol{a} = (a_1, \ldots, a_N) \in \mathcal{A}_{\text{tot}}$, called strategy *profile*, the utility of player n is denoted $u_n(\boldsymbol{a})$. In order to highlight player n in the strategy profile, common GT notation also defines $\boldsymbol{a}$ as $(a_n, \boldsymbol{a}_{-n})$, where a_n indicates the strategy of player n, while $\boldsymbol{a}_{-n}$ represents the strategies of the other players; the utility can thus be written as $u_n(a_n, \boldsymbol{a}_{-n})$. Specific functions modeling the utility in RAT selection games have been discussed in Section III-B.

In strategic games, each player chooses a strategy with no information on the strategies adopted by the others *at the present time*. This leads to considering the strategic game as an *imperfect information* game, in contrast to *perfect information* games, where the strategies adopted by the other players are known. Strategic games are usually represented in *matrix* (normal) *form*, where the utility for all players and strategy profiles is reported in a $\mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ matrix.

The assumption of imperfect information is also verified when players select their strategies in a non-simultaneous manner, but cannot observe each other. If a predetermined order among players exists, strategic games expand to *sequential* games, also known as *extensive* games. Sequential games introduce a *timeline*, and the *tree* (extensive) *form* is a common representation for these games, where a rooted tree represents players as vertices, strategies as branches, and utilities as leaves of the final branches.

Figure 5 reports an example of a RAT selection game between 2 users having 2 available RATs, in both matrix (Figure 5a) and tree (Figure 5b) forms. In the second case, User 1 applies one of its strategies (selecting RAT 1 or RAT 2) before User 2. However, the game is an imperfect information game, given that users cannot observe each other, and thus the timeline introduced by the sequential representation does not affect game dynamics and solutions. The imperfect information is represented in the tree form via a dashed line connecting the vertices that identify a player (see Figure 5b). Then, the vertices and the dashed line are globally referred to as the *information set* of the considered player. Note that, in both Fig. 5a and Fig. 5b, the utilities of the two users for each strategy profile (i.e., for all $\boldsymbol{a} = (a_1, a_2) \in \mathcal{A}_{\text{tot}}$) are reported in the form (u_1, u_2). In the example, utility is mapped onto DL throughput, and a simplified PF model is assumed (see Section III-B2); hence, users that simultaneously connect to the same RAT, equally share the RAT capacity (50 Mbps for RAT 1 and 30 Mbps for RAT 2).

In the case of extensive games with perfect information (i.e., games where a player can observe the strategies of previous players), the vertices representing a player are not connected to each other, and thus form autonomous, singleton information sets. Extensive games with perfect information have been rarely adopted in wireless communications [32]; an example for RAT selection is provided in [125].

Due to the need for learning, RAT selection is most often modeled as a dynamic game formed by multiple iterations over time. In this case, for both strategic and sequential games, each iteration is a game *step* or *stage*. Moreover, the concept of *perfect recall* is defined [31, Chapter 3]: A dynamic game is of perfect recall if, at game step t, each player knows the game *history*, i.e., the strategies applied by the other players up to step $t - 1$.

C. COMPLETE VS. INCOMPLETE INFORMATION

On the one hand, the difference between imperfect and perfect games depends on whether players observe each other when applying a strategy; on the other hand, *complete* vs. *incomplete* games differentiate with respect to the level of information each player has on the structure of the game itself, including knowledge of utility, other players, and their strategies [126]–[128].

To exemplify, the RAT selection game in Figure 5 may be considered as a complete information game when both users know they are part of a game, and also know how to represent such a game (i.e., how utility is affected by strategy profiles). Otherwise, it is an incomplete information game, where users have no knowledge of one another and may even ignore the game structure, since they may have no information on the number of other users and corresponding strategies and utilities. The hypothesis of complete information pairs with the need for learning the optimal RAT to be connected to (knowing a priori the game structure), while incomplete information also requires to learn the context (that is the

utility on each available RAT and if other users are playing the same game).

Different assumptions on complete vs. incomplete information have triggered different analyses and proposed solutions for RAT selection. As mentioned in Section IV-A, obtaining basic information on the selection game, such as the number of users and their selection strategies, is not straightforward in practice and requires signaling overhead. The trade-off between context-awareness and signaling should therefore be explicitly considered, by adopting for example network-assisted approaches, where access nodes of candidate RATs may act as selection *coordinators* beaconing relevant information to the involved users. On the other hand, context retrieval, when performed by users, may be easily affected by inaccuracies, due to measurement limitations and fast environment dynamics. Users may observe *noisy* utility samples, and thus utility should be represented as a random variable.[7] As discussed in the next sections, these aspects affect convergence and practicability of the adopted learning algorithms.

D. PURE VS. MIXED STRATEGIES

As mentioned in Section IV-B, the set of strategies $\mathcal{A}_n$ for the n-th user involved in a RAT selection game usually corresponds to the set of available RATs. These strategies are referred to as pure, and the adoption of a strategy results in selecting the RAT to connect to. Mixed strategies can be derived by defining a set of probability distributions over the elements of $\mathcal{A}_n$, denoted $\Delta(\mathcal{A}_n)$. A generic mixed strategy of the user n, denoted $\pi_n(a_n^1, \ldots, a_n^{|\mathcal{A}_n|}) \in \Delta(\mathcal{A}_n)$, assigns a probability of selection to each available RAT.[8] That is, a probability $p_n(a_n)$ is assigned to each pure strategy in $\mathcal{A}_n$, such that $\pi_n(a_n^1, \ldots, a_n^{|\mathcal{A}_n|}) := (p_n(a_n^1), \ldots, p_n(a_n^{|\mathcal{A}_n|}))$. It thus follows that a pure strategy is a mixed strategy where the probability associated to the *purely* selected RAT is 1.

A joint probability distribution defined over $\Delta(\mathcal{A}_{\text{tot}})$, referred to as mixed strategy profile, can be obtained by multiplying all marginal distributions, thus assuming independence among each player selection. It is indicated as $\boldsymbol{\pi} = (\pi_1(a_1^1, \ldots, a_1^{|\mathcal{A}_1|}), \ldots, \pi_N(a_N^1, \ldots, a_N^{|\mathcal{A}_N|}))$ or simply $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. GT notation $(\pi_n, \boldsymbol{\pi}_{-n})$ is often used to emphasize the contribution of a generic user n. A more general definition relaxes the assumption of independence between players. In this case, the joint probability distribution is denoted $\boldsymbol{\Pi}$. In the context of RAT selection, this models situations where users actuate their selections by following *cooperation policies*, either injected by selection coordinators or autonomously derived through learning.

[7] It can be observed that noisy information may also be considered for other game parameters rather than utility, e.g., number of available RATs and users. To the best of our knowledge, however, such a modeling option is rarely adopted in RAT selection. In most cases, these information are assumed to be either known (with no uncertainty) or unknown.

[8] $\pi_n(a_n^1, \ldots, a_n^{|\mathcal{A}_n|})$ is a distribution over the elements in $\mathcal{A}_n$. Hence, it is a vector of probabilities $p_n(a_n)$, for all $a_n \in \mathcal{A}_n$, such that $\sum_{a_n \in \mathcal{A}_n} p_n(a_n) = 1$. It is not denoted using a bold notation since this is used to represent a non-singleton group of players.

Both pure and mixed strategies have been adopted in RAT selection games. Mixed strategies are, however, more general, and better adapt to the dynamic nature of these scenarios. As discussed in Section IV-E, the adoption of pure vs. mixed strategies determines the type of solution (i.e., the type of equilibrium) that can be achieved and, in turn, the learning algorithm to use (Section VI).

E. SOLUTION CONCEPTS

The solution of a non-cooperative game is the equilibrium. Pure and mixed Nash equilibrium (NE) and correlated equilibrium (CE) are widely considered in strategic games. The definitions provided below naturally map to complete information games. However, the same concepts can be adopted in incomplete information games since equilibria can be achieved by using learning algorithms.

Definition 1 (Pure Nash Equilibrium (PNE) [114]): A pure strategy profile $(a_n^{\mathsf{NE}}, \boldsymbol{a}_{-n}^{\mathsf{NE}})$ is a PNE if and only if it satisfies the following condition:

$$u_n(a_n^{\mathsf{NE}}, \boldsymbol{a}_{-n}^{\mathsf{NE}}) \geq u_n(a_n, \boldsymbol{a}_{-n}^{\mathsf{NE}}) \quad \text{for all } n \in \mathcal{N} \text{ and } a_n \in \mathcal{A}_n. \tag{3}$$

PNEs generalize into MNEs if mixed strategies are used.

Definition 2 (Mixed Nash Equilibrium (MNE) [114]): A mixed strategy profile $(\pi_n^{\mathsf{NE}}, \boldsymbol{\pi}_{-n}^{\mathsf{NE}})$ is a MNE if and only if it satisfies the following condition:

$$\hat{u}_n(\pi_n^{\mathsf{NE}}, \boldsymbol{\pi}_{-n}^{\mathsf{NE}}) \geq \hat{u}_n(\pi_n, \boldsymbol{\pi}_{-n}^{\mathsf{NE}}) \quad \text{for all } n \in \mathcal{N} \text{ and } \pi_n \in \Delta(\mathcal{A}_n). \tag{4}$$

$\hat{u}_n(\pi_n, \boldsymbol{\pi}_{-n})$ represents the *expected* utility of the n-th player when the mixed strategy profile $(\pi_n, \boldsymbol{\pi}_{-n})$ is played, that is:

$$\hat{u}_n(\pi_n, \boldsymbol{\pi}_{-n}) := \sum_{\boldsymbol{a} \in \mathcal{A}} \left[\prod_{n' \in \mathcal{N}} p_{n'}(a_{n'}) \right] u_n(\boldsymbol{a}). \tag{5}$$

When the assumption of independence among players' mixed strategies is relaxed, MNEs are found to be CEs.

Definition 3 (Correlated Equilibrium (CE) [129]): A joint strategy profile $\boldsymbol{\Pi}^{\mathsf{CE}}$ is a CE if and only if it satisfies the following condition:

$$\sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\Pi}^{\mathsf{CE}}(a_n, \boldsymbol{a}_{-n}) u_n(a_n, \boldsymbol{a}_{-n}) \geq \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\Pi}^{\mathsf{CE}}(a_n, \boldsymbol{a}_{-n}) u_n(\mathsf{f}_n(a_n), \boldsymbol{a}_{-n}) \quad \text{for all } n \in \mathcal{N} \text{ and functions } \mathsf{f}_n : \mathcal{A}_n \to \mathcal{A}_n, \tag{6}$$

where f_n is any function allowing the n-th player to unilaterally change its played pure strategy.

By definition, once in a PNE, MNE, or CE profile, no player has an interest in changing strategy, given that no utility gain is obtained by doing so under the assumption that the others do not change their own strategies either. At the equilibrium, each player plays a so-called *best response* (BR) strategy, that is, a strategy that maximizes utility in response

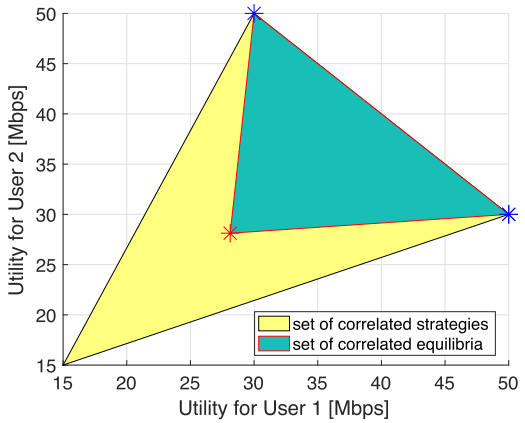

FIGURE 6. Solution space for the RAT selection game of Figure 5, as a function of utility of users. The set of possible CEs (green area) is shown against the entire set of correlated strategies of users (yellow area). Blue and red stars highlight the two PNEs and the single MNE for the game, respectively.

to (i.e., *given*) the strategies applied by the others. Multiple BRs may exist for a player (not all leading to equilibria); in the following, given a strategy profile $\boldsymbol{a}_{-n}$, the set of possible BRs of user n given $\boldsymbol{a}_{-n}$ is denoted $\{\mathsf{BR}_n\}|\boldsymbol{a}_{-n}$.

Figure 6 shows the solution space for the RAT selection game in Figure 5. The game has two PNEs, i.e., users interchangeably split over the two available RATs. In these two cases, the user connecting to RAT 1 experiences a utility (throughput) of 50 Mbps, while the other user has a throughput of 30 Mbps, on RAT 2. The game also possesses a MNE, under which both users select RAT 1 and RAT 2 with probabilities equal to 0.875 and 0.125, ultimately experiencing the same average throughput (slightly above 28 Mbps). Figure 6 shows that such NEs are the corner points (blue and red stars) of a wider set of solutions, i.e., the set of CEs (green area), at which the game can converge without the assumption of users independence.

Also note that both users do not hold a *strictly dominant* strategy, i.e., a strategy leading to the highest utility, irrespectively of the strategies of the others.[9] Furthermore, the two PNEs of this game are Pareto-optimal (PO) profiles, i.e., they lead to the highest utility for both users. Such profiles are also social-optimum (SO) points, since the sum of users utilities is the highest compared against the sum of other profiles. The utility comparison between PO, SO, and NE profiles gives initial indications on the optimality of the game equilibria.

GT provides several extensions to the above definitions. In particular, ε-equilibria can be formalized. They somehow relax the above definitions and make convergence simpler. In Definition 1–3, each player keeps the current strategy since it would experience a variation in utility $\Delta u_n \leq 0$ if it selected any other strategy. In an ε-equilibrium, a player keeps the current strategy since it would observe a $\Delta u_n \leq \varepsilon$ for any other strategy. An application to RAT selection is given in [102].

Moreover, considering extensive games, a NE refinement named subgame perfect (Nash) equilibrium (SPE), is also used, given the definition of subgame [130]. Conceptually, while some NEs may be somehow counter intuitive with respect to rational thinking, SPEs represent *credible* solutions among all possible equilibria [59].

Table 2 summarizes the concepts discussed in this section, by highlighting how the analyzed GT modeling options map onto RAT selection scenarios, and provides a handy reference toward setting up a RAT selection game.

V. RAT SELECTION GAMES

In the previous section, we have reviewed game-theoretic modeling choices for RAT selection, in terms of adopted settings and practical considerations. This section complements the former by describing three non-cooperative games, that is, potential (and congestion) games, evolutionary games, and stochastic games, all of which have been applied to RAT selection, in particular in their dynamic form. A discussion on how the models match with practical scenarios is also provided. It is worth mentioning that two further non-cooperative models, known as Bayesian and Stackelberg (or Leader-Follower) games, have been used to model RAT selection. While Bayesian models have found limited application to RAT selection (an example can be found in [95]), Stackelberg games require networks (RATs) to also take actions and observe the corresponding utility [131]–[133]. In the following, we will neither discuss Bayesian games, due to their limited application, nor Stackelberg games, since the focus of this paper is on strategies involving users vs. users competition.

A. POTENTIAL AND CONGESTION GAMES

One of the most adopted models for RAT selection is the potential game, in particular in its congestion form [134], [135]. This is due to the properties of such games in terms of utility and existence of PNEs, as described below.

Definition 4 (Potential Game (PG)): A strategic game $\mathcal{G}$ is a PG if and only if there exists a global function $\mathsf{F} : \mathcal{A}_{\text{tot}} \rightarrow \mathbb{R}$, common to all players and referred to as *potential*, that can express the change in utility observed by players when they change strategies.

The type of relationship between F and utility drives a taxonomy across PGs [135]. In particular, *exact* PGs can be defined as follows:

Definition 5 (Exact Potential Game (EPG)): A PG is an EPG if and only if, for each player, the difference between the utility of two strategies (given all other strategies being equal) results in the same difference in the potential function, i.e.:

$$u_n(a_n, \boldsymbol{a}_{-n}) - u_n(a'_n, \boldsymbol{a}_{-n}) = \mathsf{F}(a_n, \boldsymbol{a}_{-n}) - \mathsf{F}(a'_n, \boldsymbol{a}_{-n}) \quad \text{for all } n \in \mathcal{N},\ a_n, a'_n \in \mathcal{A}_n \text{ and } \boldsymbol{a}_{-n} \in \mathcal{A}_{\text{tot}} \setminus \mathcal{A}_n. \quad (7)$$

[9] If a strictly dominant strategy exists, it is played by a player in a NE. Consequently, strategy profiles including *strictly dominated* strategies for any particular player can be discarded while looking for NEs.

TABLE 2. GT modeling options and their use in user-centric RAT selection.

Modeling option	Use in RAT selection
Cooperative / Non-cooperative games	Both are used but non-cooperative is predominant since it avoids information exchange between users. Possible limitations may be addressed via network-assisted mechanisms.
Strategic / Extensive forms	Strategic form is predominant since it avoids setting a selection order among users. It also implies imperfect information, i.e., users cannot observe the selection of the others at the present time.
Complete / Incomplete information	Both are used. Complete information implies that users know the utility and how it is affected by the others; Incomplete information requires the observation of samples of the utility, whose definition may be unknown. In this case, possible limitations may be addressed via network-assisted mechanisms.
Pure / Mixed strategies	Both are used, depending on the adopted game model (e.g., existence of PNEs or MNEs) and learning scheme (e.g., the algorithm may imply the type of strategy to be used, as discussed in Section VI).
Solution concepts	Several types of equilibrium are used, depending on the adopted game model and learning scheme (e.g., the use of an algorithm may imply which type of equilibrium is achievable, as discussed in Section VI).

The following theorem holds for EPGs and other PGs:

Theorem 1: A PG with a finite number of players, each one having a finite number of possible strategies, i.e., a *finite* PG, always admits at least one PNE.[10]

Moreover, a correspondence between EPGs and so-called congestion games exists [134], [135].

Definition 6 (Congestion Game (CG)): A strategic game $\mathcal{G}$ is called CG if and only if, for each player, the utility of a given strategy is a monotonically non-increasing function of the number of players adopting that strategy.

Note how the RAT selection game in Figure 5 follows Definition 6. CGs are most often defined in terms of cost, rather than utility, associated with the available strategies; then, by thinking strategies as *resources*, it follows that the congestion cost (or *load*) experienced when selecting a resource is a strictly non-decreasing function of the number of players selecting the resource. The cost function may be different among players along with pure strategy sets $\mathcal{A}_1, \ldots, \mathcal{A}_N$; in the latter case the CG is asymmetric, in contrast to a symmetric CG, which is also referred to as *crowding* game.

The aforementioned correspondence between EPGs and CGs is always verified in the case of *unweighted* congestion games (uCGs), where the players congest the resources equally, in contrast to *weighted* congestion games (wCGs), where the players have different weights related to their contribution in congesting a resource. In uCGs, an exact potential function referred to as Rosenthal's potential can be expressed [112], [134], meaning that the game is also an EPG. Corollary 1 follows from Theorem 1:

Corollary 1: uCGs always admit at least one PNE.

As discussed in Section III-B, a definition of cost (utility) depending on the number of users selecting a RAT, e.g., the interference level or the achieved throughput, expresses the rationale of using a CG to model user-centric RAT selection [94], [102], [103], [108], [112], [136]. Hence, Corollary 1 has been used in initial work on GT modeling of RAT selection to demonstrate the existence of PNEs. Moreover, two further aspects reinforce this modeling choice:

- the existence of PNEs for wCGs, that in general do not admit potential functions, has been demonstrated under some constraints on the cost function [137], [138];
- the possibility to converge to PNEs in both uCGs and wCGs by adopting several distributed learning algorithms was proved [32], [33] (see Section VI).

Application of Potential and Congestion Games to RAT Selection: PGs and CGs were adopted in initial work on game-theoretic modeling of RAT selection, thus triggering an initial prevailing focus on PNEs. A cost function dependent on the number of users connected to each RAT, but independent of exogenous factors related to context (e.g., radio conditions), allowed the analysis of theoretical bounds on the efficiency of distributed selection schemes converging to PNEs [112]. Under the assumption of complete information, full rationality, and observability of other users' selection strategies, the aspect of learning the equilibria was analyzed in [102], where the BRD algorithm was used to achieve convergence to PNEs (see Section VI-A1). The model was also extended to a multi-leader / multi-follower game, where user competition for selecting the best RAT was anticipated by RAT competition for selecting the best frequency resource. In this case, convergence to ε-SPEs was shown.

A CG model was also given in [108], where the unique PNE of the proposed game was achieved via BRD under complete information, and via a RL scheme based on Q-Learning in the incomplete information case (see Section VI-B).

PGs and CGs provide a specific structure for RAT selection games in terms of utility (cost), albeit not directly embedding the aspects of dynamicity and learning proper of such scenarios. The introduction of advanced learning schemes under a more realistic assumption of incomplete information has driven the interest on broader solution concepts, such as MNEs and CEs, thus relaxing the need for a PG or CG model. On the one hand, this paradigm shift led to more complicated analyses, due to the impossibility of evaluating equilibria through the study of a player-independent function[11]; on the other hand, it has allowed to extend the analysis,

[10] Theorem 1 relies on the *acyclicity* property of PGs, that allows to define *finite improvement paths* (FIPs). End points of FIPs are PNEs [135].

[11] PNEs corresponds to local and global minima of the potential cost function. Hence, a study of the potential function reveals the equilibria.

by considering aspects not directly associated to the level of congestion of the selected RAT but still affecting utility, e.g., billing costs, energy efficiency, and mobility.

B. EVOLUTIONARY GAMES

Another approach for RAT selection modeling is evolutionary game theory (eGT), that takes inspiration from biology and Darwinian evolution [58, Chapter 6] [139], [140]. eGT extends non-cooperative GT by introducing the concept of *population*. This concept can be introduced in RAT selection as well, where one or different populations may implicitly refer to groups of users having a same set of candidate RATs, based on their position and the coverage area of the access nodes of each RAT [96].

eGT also relaxes the assumption of full rationality, that considers players always eager to maximize their utility, and favors the hypothesis of bounded rationality, that assumes players to be cognition-limited and thus eager to learn over time[12] [139], [141]. Once again, the assumption nicely maps onto RAT selection situations, where users often have limited knowledge on utility and other features.

In general, an evolutionary game is a repeated interaction between players belonging to a same or different populations. For the sake of simplicity, a single-population game is assumed from now on. At each t-th game step, $N_{\mathcal{P}}$ players forming a population $\mathcal{P}$ apply one of the available pure strategies in the population strategy set $\mathcal{A}_{\mathcal{P}}$.[13] Let us define the proportion of usage for the generic strategy $a_{\mathcal{P}} \in \mathcal{A}_{\mathcal{P}}$ as $f_{a^i_{\mathcal{P}},t} := \frac{N_{a^i_{\mathcal{P}},t}}{N_{\mathcal{P}}}$, where $N_{a^i_{\mathcal{P}},t}$ is the number of players adopting strategy $a^i_{\mathcal{P}}$ at time t. The *population state* $f_{\mathcal{P},t}$ is then defined as the set of strategy proportions at time t, that is, $f_{\mathcal{P},t} := (f_{a^1_{\mathcal{P}},t}, \ldots, f_{a^{|\mathcal{A}_{\mathcal{P}}|}_{\mathcal{P}},t})$.

As detailed in [31, Chapter 3], two interpretations can be considered for the analysis of evolutionary games. In the more general setting, the utility of the adopted strategy $a^i_{\mathcal{P}}$ at time t depends on the population state at that same time, and is referred to as $u(a^i_{\mathcal{P}}, f_{\mathcal{P},t})$. It follows that the average population utility is:

$$\overline{u}(f_{\mathcal{P},t}) = \sum_{a_{\mathcal{P}} \in \mathcal{A}_{\mathcal{P}}} f_{a_{\mathcal{P}},t} u(a_{\mathcal{P}}, f_{\mathcal{P},t}). \tag{8}$$

The goal of eGT is to analyze how an evolutionary process drives a change of the population state toward a *stable* population composition, i.e., an evolutionary equilibrium. An evolutionary process promotes the *selection* of strategies that perform better than others, thus triggering an increase of proportions over time, but also a *mutation* of strategies, that leads to new population states. The balance between selection and mutation leads to the definition of several evolutionary processes. Replicator dynamics (RD), that focuses on the aspect of selection, is the most adopted process [142]. In the analysis of a single-population game via RD, the *rate* of change of a strategy proportion is regulated by the difference between the utility obtained by the players adopting that strategy and the average utility of the population. In continuous time, this corresponds to a set of ordinary differential equations. For strategy $a^i_{\mathcal{P}}$, one has:

$$\frac{\partial f_{a^i_{\mathcal{P}},t}}{\partial t} = f_{a^i_{\mathcal{P}},t}[u(a^i_{\mathcal{P}}, f_{\mathcal{P},t}) - \overline{u}(f_{\mathcal{P},t})]. \tag{9}$$

Players adopting strategies leading to a utility higher than the average population utility *replicate* themselves faster than others, and thus their presence in the population grows. An equilibrium is found by evaluating the steady-state of the system of differential equations, that is, by equating to zero and solving the replicator dynamics in (9), for all $a_{\mathcal{P}} \in \mathcal{A}_{\mathcal{P}}$. Several theoretical results justify the use of RD for the analysis of evolutionary games [27, Chapter 5] [28], in particular:

Theorem 2: Any pure or mixed strategy profile that is a PNE or MNE of the evolutionary game is a steady-state of the RD.

Theorem 2 highlights that RD may converge to NEs. A stronger condition can be derived by considering another solution concept referred to as *evolutionary stable strategy* (ESS). ESS is a refinement of a NE ($\{\text{ESSs}\} \subset \{\text{NEs}\}$) and takes into account the aspect of mutation [140]. From the evolutionary idea of *survival of the fittest*, ESS is a strategy configuration that, being played by the population, overcomes in utility any other configuration obtained as a result of the mutation (i.e., the selection of a different strategy) of an arbitrary small proportion of players. The following theorem holds for ESSs [142]:

Theorem 3: An ESS is an asymptotically stable steady-state of the RD.

Theorem 3 highlights that RD may converge to ESSs under the constraint of *asymptotic stability* of the RD solution, that implies Lyapunov stability [143]. Theorems 2 and 3 provide interesting insights on RD and its relation with NEs and ESSs. However, they do not provide a practical way of finding such solutions. RD does not always converge to NEs or ESSs, and the reciprocal of both theorems is not valid in a general sense. However, the following theorem holds [28]:

Theorem 4: An asymptotically stable steady-state of the RD is a NE of the evolutionary game.

The above theorems show that the convergence of RD to an asymptotically stable steady-state corresponds to the convergence of the evolutionary game to a NE, that may also be an ESS. In order to confirm the game convergence to a NE, Theorem 4 indicates to check the asymptotic stability of the RD solution, which is verified when all the eigenvalues of the Jacobian matrix associated to the RD have a negative real part [144].

For example, Figure 7 shows the convergence of RD to the MNE of the RAT selection game in Figure 5, independently of the initial population state. As a matter of fact, such a game

[12] The interested reader may refer to [141], and references therein, for an insightful discussion on how the word *evolution* and the learning process in eGT can be interpreted in either biological or cultural sense.

[13] In this section, the subscript assigned to the game parameters stands for the entire population of players. The population is *homogeneous* if the strategy set $\mathcal{A}_{\mathcal{P}}$ unconditionally applies to all players.

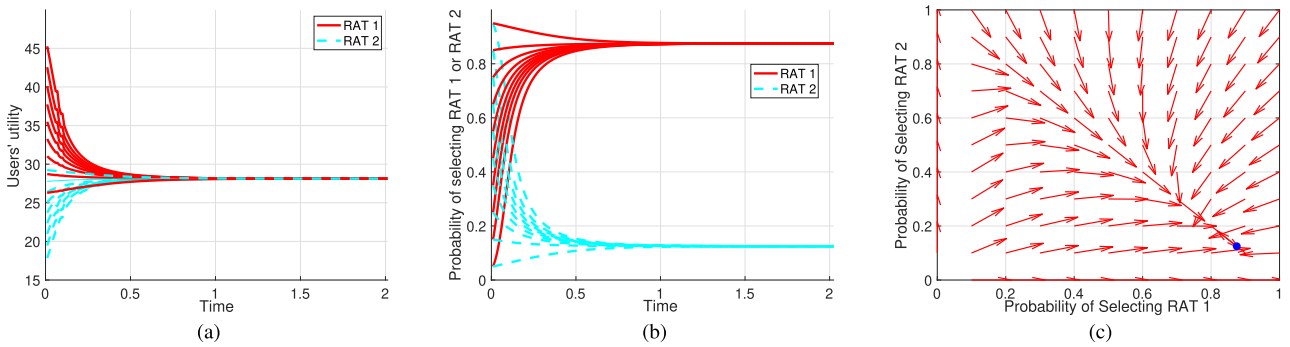

FIGURE 7. Application of RD to the RAT selection game of Figure 5. Convergence to the MNE is reported in terms of users utilities (a) and probability of selecting RAT 1 (red lines) and RAT 2 (light-blue dashed lines) actions (b), for different initial population states. The complete vector field of probabilities of selecting RAT 1 vs. RAT 2 and the convergence point (blue dot) are given in (c).

can be thought as a single-population evolutionary game due to the symmetry of the utility matrix.

Application of Evolutionary Games to RAT Selection: Several reasons have triggered the use of evolutionary games in RAT selection, including the possible adoption of the concept of population and of the hypothesis of bounded rationality. Moreover, RD provides a powerful method for studying game dynamics. Indeed, it is usually considered as a benchmark for the convergence of distributed RAT selection schemes, since finding the RD solution requires to observe the decision and the utility of each user, in order to compute the average population utility. As derived in Theorems 2–4, RD convergence to an asymptotically stable steady-state provides a sort of refined solution in terms of NE stability, and thus may be used to compute specific equilibria in games with multiple equilibria, and to check if distributed schemes lead to same or different solutions.

A RD-inspired network-assisted selection scheme was proposed in [96] and referred to as *population evolution*. The RD solution was adopted as a benchmark for both this scheme and a fully distributed algorithm based on Q-Learning (see Section VI). Q-Learning convergence was also compared against RD in [108]. The model in [96] was reused in [145], adding further network constraints. It was assumed that network operators may want to set some limits on their RATs, e.g., the maximum number of accepted users, in order to maintain the QoS to predefined values. This aspect was analyzed by adding a constraint to one out of three available RATs. The constraint was an indication function added to the definition of utility for that particular RAT, triggering a sharp utility decrease when the constraint is violated (*punishment*). Such a game was finally solved via the population evolution algorithm proposed in [96]. Simulation results showed that the punishment term biases the equilibrium: users stop choosing that RAT when the game reaches the step where the constraint was violated. This resulted in the termination of the competition for that RAT, that instead continued for the other candidates until an equilibrium was achieved for them as well. The approach allowed to maintain the QoS of the first RAT above a predefined minimum threshold. A similar analysis was presented in [146], where, however, the utility function was represented by a weighted combination of sub-utilities obtained from different network attributes. Results showed a rather fast convergence to the set of equilibria via a RD-based scheme. In [145], [146], the IEEE 802.21 Media Independent Handover (MIH) protocol [147] was assumed to provide the information on the other users in the same area, so that each user could apply the algorithm based on RD. The work in [146] was extended in [148], that considered different users, different traffic (while [146] only considered streaming users), and 4G and 5G RATs.

From these works, two limitations in using eGT and RD for RAT selection emerge. First, the evolutionary model requires a player-independent utility, i.e., all players in the population that apply the same strategy have the same utility. This limits the available utility functions. Secondly, RD drives the system toward a fair configuration: at equilibrium, players do not change strategies since they achieve a utility equal to the average population utility. RD thus promotes load balancing across RATs, which may be desirable but less efficient compared to other solutions (note in Figure 7 how the RD applied to the RAT selection game of Figure 5 converges to the fairest NE among the available ones, where users achieve the same utility).

Both issues are recently addressed in [149], that proposes the use of *fractional* eGT in a RAT selection scenario involving macrocells and mmWave small cells, some of it being provided by unmanned aerial vehicles (UAVs). Differently from classic eGT, fractional eGT includes memory of players in learning dynamics. Hence, users take into account instantaneous utility and previous decisions in the selection process [150]. In particular, the power-law memory is used, since it is widely adopted in economic processes and is experimentally validated against the behaviour of human memory.[14] The memory effect is incorporated in RD by replacing the integer-order time derivative with the left-sided Caputo fractional derivative of order β [150]. The case $\beta = 1$ represents classic eGT; hence, $\beta \in (0, 1)$ and $\beta \in (1, 2)$

[14]See [149] for more details.

are also analyzed. The scenario with user-independent utility is extended to situations where users have a different utility (i.e., a different throughput) depending on radio conditions. In this setup, existence, uniqueness, and stability of a fractional evolutionary equilibrium are theoretically derived and numerically demonstrated. A positive memory effect is obtained when $\beta > 1$, leading RD to converge to a fractional equilibrium having higher per user utility than its classic counterpart.

C. STOCHASTIC GAMES

User-centric RAT selection is also modeled as a stochastic game, also known as Markov game, that can handle the hypothesis of incomplete information [151]. With respect to the strategic game $\mathcal{G}$, the stochastic game $\mathcal{G}^{\mathsf{S}}$ introduces the concepts of *state* and transition probabilities between different states. A state can be thought of as an indication of the conditions of the *environment* the players face during the game. The environment and its dynamics are partially or fully unknown when incomplete information is assumed. The strategies selected by the players lead the environment to transit across states. The states may have, in turn, an impact on the utility experienced by players.

A stochastic game is formally represented by the tuple $\mathcal{G}^{\mathsf{S}} := \{\mathcal{N}, \mathcal{S}, \Delta(\mathcal{S}), \{\mathcal{A}_n\}_{n\in\mathcal{N}}, \{u_n\}_{n\in\mathcal{N}}\}$, where $\mathcal{S}$ represents the state space with elements $s \in \mathcal{S}$. $\Delta(\mathcal{S})$ is a probability distribution over $\mathcal{S}$, containing the probabilities $p(s'|s, \boldsymbol{a})$ of switching from state s to state s' given that a strategy profile $\boldsymbol{a}$ is applied by players. $\Delta(\mathcal{S})$ is Markovian, since the next state s' only depends on $\boldsymbol{a}$ and the previous state s.[15] $\Delta(\mathcal{S})$ is also stationary, since it does not change over time.[16] The definition given for $\mathcal{G}^{\mathsf{S}}$ assumes a unique state space across players and state-independent pure strategy sets for each player. However, player-specific states and state-depending strategy sets may be considered for more general models [31, Chapter 3]. In all cases, players utility at each game step depends on both states and strategies.

The dependence of the game parameters on states suggests a practical interpretation of a stochastic game; it can be seen as *a repeated interaction among players in a game with changing structure over time*. As shown in Figure 8, the fundamental caveat of this interpretation is that these games can be seen as MDP and MAB extensions to multi-agent

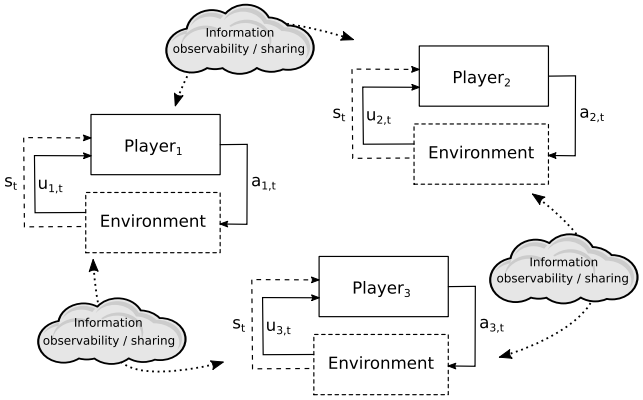

FIGURE 8. Representation of a stochastic game. Dashed lines for the environment and states depict the possibility of observing or not any of them. Information observability and sharing across players (clouds with dashed lines) regulate perfect/imperfect and complete/incomplete information assumptions.

scenarios. This mapping is further discussed in Sections VI-B and VI-C, where more detail on MDP and MAB models is given prior to presenting the RL-based solutions applied to RAT selection. To exemplify, assuming a probabilistic observability of state at time t, i.e., s_t, for each player and for all t, the stochastic game can be seen as a multi-agent extension of a partially observable MDP (POMDP). Moreover, stochastic games may define player-dependent states, i.e., s_t is replaced by $s_{n,t}$, for all $n \in \mathcal{N}$. Similarly, if states also depend on actions, i.e., s_t is replaced by $s_{a_{n,t},t}$, for all $a_n \in \mathcal{A}_n$ and $n \in \mathcal{N}$, then the stochastic game represents a multi-agent extension of a stateful MAB. On the contrary, with no definition of states, the stochastic game is a multi-agent extension of a stateless MAB.

In all cases, the goal of players in $\mathcal{G}^{\mathsf{S}}$ is similar to the goal of a single agent in MDP and MAB, that is, to find the strategy at each state and time that maximizes their utility. Hence, stochastic games inherit MDP and MAB tools in addressing aspects such as learning and adaptation to an unknown environment [27, Chapter 1]. During the game, players thus learn and build a strategy plan over the state (and time) space. Such a plan is generally referred to as *policy* in the MDP and MAB literature. Considering pure strategies, the generic pure policy for player n can be denoted as $a_n^{\mathsf{P}} = \{a_{n,t}\}_{t\geq t_0}$, where $a_{n,t}$ identifies the strategy applied at time t, and t_0 is the time at which the policy starts to be adopted by the player. The dependence on the state is implied since a state is visited at each time t. A pure policy profile can then be written as $\boldsymbol{a}^{\mathsf{P}} = (a_1^{\mathsf{P}}, \ldots, a_N^{\mathsf{P}})$ or $(a_n^{\mathsf{P}}, \boldsymbol{a}_{-n}^{\mathsf{P}})$. Therefore, the goal becomes the assessment of a policy profile, i.e., a strategy profile for each state (and time) that maximizes the expected utility of each player. The latter is evaluated by cumulating the instantaneous utilities experienced from t_0 (and corresponding state) on, meaning that the final goal is to maximize a *long-term* expected utility. Under the commonly adopted *infinite time-horizon discounted model*, that is, the game extends over $t \to +\infty$, the generic player n

[15]The Markovian assumption is valid from a game-wide perspective, i.e., by considering the entire strategy profile $\boldsymbol{a}$ and its relation with state transitions. However, it is not valid from the perspective of players applying their strategies without observing the others, i.e., *independent learners* in imperfect and incomplete information games, since they do not have a game-wide perspective. This situation arises when SARL-native schemes, e.g., Q-Learning, are adopted in multi-agent scenarios (see Section VI-B) [152].

[16]Similarly to the Markovian assumption, multi-agent scenarios pose a challenge with respect to the environment stationarity. The assumption holds from a game-wide perspective, since it allows to differentiate between environment and players. However, independent learners cannot discriminate between environment and other agents. Hence, the other agents are part of the environment and contribute to its non-stationarity due to their learning process [153].

tries to maximize its so-called expected *discounted utility*, also referred to as *return*, defined as follows:

$$U_{n,t_0} := \mathbb{E}\left[\sum_{t=0}^{+\infty} \lambda^t u_{n,t_0+t+1}\right], \tag{10}$$

where, for the sake of simplicity, a notation more common to MDP and MAB rather than GT is adopted (note that the dependence of u_n on actions of players, e.g., a_n and $\boldsymbol{a}_{-n}$, is omitted). U_{n,t_0} is the return for player n from time t_0 on, and $\mathbb{E}[\cdot]$ is the expectation over the environment state dynamics, the (possibly mixed) policy profile, and the utility function.[17] Moreover, $u_{n,t}$ denotes the instantaneous utility of player n (for all $t > t_0$), and λ ($0 < \lambda < 1$) is the so-called *discount factor*, that can be interpreted as the interest of players in short vs. long-term returns.[18]

While trying to maximize their return, players possibly converge to equilibria, the definitions of which closely follow the ones provided in Section IV-E, where the concept of policy profile is substituted to the one of strategy profile.

The above discussion considers pure policies. Given an infinite stochastic game, a pure policy is an infinite sequence of pure strategies. This challenges the definition of mixed policies and thus another type of policy, referred to as *behavioral*, becomes relevant [31, Chapter 6]. A behavioral policy allows a player to randomize its pure strategies in $\mathcal{A}_n$ at each game step (i.e., at each state).[19] The Kuhn's Theorem [154] allows to restrict the game analysis on behavioral policies, since there exists a correspondence between behavioral and mixed policies under the assumption of either perfect information or perfect recall [31, Chapter 3]. The existence of equilibria in terms of *stationary* behavioral policies, i.e., policies where the association between strategies and states does not change over time, was proven for stochastic games with finite sets of states and pure strategies [155].

Application of Stochastic Games to RAT Selection: The adaptation of the above model to RAT selection requires several observations. First, the definition of states depends on the adopted model. For example, in the single-user MDP model for vertical handover given in [156], the state at time t is a joint information of available bandwidth and experienced delay. RATs share information with users, that in turn use it to evaluate the transition probabilities and an optimal, stationary, pure policy, i.e, a RAT to connect to given a state. Furthermore, the state observability is not certain and may require network assistance [156]. Hence, POMDPs, that generalize MDPs assuming a probabilistic state observation, have also been proposed [157]. In general, modeling RAT selection as a stochastic game does not even require an explicit definition of states, as shown in [96], that is analyzed in detail in Section VI-B1.a. From this perspective, also note that the RAT selection game of Figure 5 can be thought of as a simplified stochastic game (with no definition of states and a deterministic utility).

Secondly, once states are defined, the knowledge on how the environment transits across states and how states affect utility should be carefully discussed. From a single-agent perspective, assuming to know the state transition probabilities and the utility function allows using dynamic and linear programming policy discovery schemes, such as Value and Policy Iteration algorithms [39, Chapter 4]. For example, Value Iteration is used in [156] to derive the optimal RAT for each state. However, transition probabilities and utility may be unknown while looking for candidate RATs. Hence, RL schemes could be used for discovering the context and find an optimal policy. This aspect is more challenging in multi-agent scenarios, considering that the learning process of agents enhances the dynamicity of the system under analysis. Such a situation naturally maps onto a stochastic game with incomplete information [31, Chapter 6], the solution of which may still rely on RL. However, the use of RL algorithms is not straightforward and depends on the assumptions on rationality of players and game observability. Moreover, it is not guaranteed that RL algorithms converging to optimal policies in single-agent setups also converge to optimal policies in multi-agent scenarios (Section VI).

In a RAT selection game, the focus is not on perfectly predicting which equilibrium is achieved, but rather to address the challenges related to the discovery of a policy driving the selection process toward an equilibrium. Hence, it is key to compare different learning schemes in terms of achievable utility, required information, convergence speed, amount of RAT switchings, etc. Such a comparison is particularly relevant under a stochastic incomplete information game model, since in this case users can adopt different RL algorithms, that may differ in terms of the above indicators.

A stochastic game model reinforces the trade-off between solution efficiency, convergence speed, and practicability of the adopted learning scheme. For example, algorithms reaching less-efficient utility solutions in a short time may be preferred to more rewarding but slower schemes, since a possibly high scenario variability due to mobility, arrival and departure of users, and other factors, might quickly nullify the achieved convergence and require a new learning process. Moreover, such a process may require users to switch several times from one RAT to another before reaching a stable policy profile, which is energy-inefficient; it then follows that the design of learning schemes that require a few switchings is

[17] The instantaneous utility adopted in MDP, MAB, and thus Markov games is in general stochastic. For example, in so-called finite MDPs, strategies and states are finite and the utility for each (state, strategy) pair has discrete outcomes with given probabilities [39]. In RAT selection, this aspect usually maps onto the concept of *noisy* utility (see Section III-B).

[18] The discount factor is assumed equal across players. Moreover, U_{n,t_0} is in its non-normalized version, which is widely adopted in MDPs. Normalized returns have a weighting term $\lambda(1-\lambda)^{t-1}$ [31, Chapter 3].

[19] To exemplify, consider a one-stage extensive game between two players, where both players play one strategy each, and then the first player can play one more time. In this case, a mixed strategy for the first player corresponds to randomize its two actions jointly, i.e., over the entire game tree. A behavioral strategy allows instead to randomize *step-by-step* independently, i.e., over each information set forming the game tree. In the example, the first player adopts a behavioral strategy if it randomizes its first action and then the other, independently. Behavioral policies extend behavioral strategies over the state set.

TABLE 3. Applicability of non-cooperative games to RAT selection.

Model	Utility Definition	Information Assumption	Learning[a]	Literature Examples	Solution
Potential and Congestion game	Dependent on the number of users selecting the same RAT	(Im)perfect and Complete	BRD	[44] [102] [108]	PNEs
		Perfect recall and Complete	FP		
		Imperfect and Incomplete	RL-based		
Evolutionary game	User-independent (users selecting the same RAT have the same utility)	Perfect and Complete	RD	[96] [149]	NEs, ESS
		Imperfect and Incomplete	RL-based		
Stochastic game	Dependent on the environment state (e.g., user and/or network indicators)	Perfect and Complete	Value (or Policy) Iteration	[156] (single-agent) [108] [96] [93]	NEs, CEs, SPEs
		Imperfect and Incomplete	RL-based		

[a] The algorithms are described in Section VI.

key for RAT selection. Due to the strong connections with learning schemes, literature examples of RAT selection modeled as stochastic games are discussed in Section VI, where the algorithms are also introduced.

Table 3 summarizes the concepts discussed in this section, highlighting how the analyzed games can be used for modeling user-centric RAT selection, in terms of utility definition, information observability, learning schemes (introduced with literature examples in Section VI), and achievable solutions.

VI. MULTI-AGENT LEARNING FOR RAT SELECTION

This section reviews learning algorithms commonly adopted for solving user-centric RAT selection. The main results in terms of game-theoretic convergence are also discussed. The applicability of the algorithms depends on how the game is formulated. As a matter of fact, BRD and FP are prominent examples of fully rational, learning-the-equilibria mechanisms; hence, their use in RAT selection requires extensive control message exchange across network entities. RL algorithms, however, can be applied when the selection problem is modeled as an incomplete and imperfect information game, which implies bounded rationality of users. Enhanced algorithms can be adopted if information sharing is assumed to take place during the selection process, and the observed information can be used together with utility.

A. GT-BASED SOLUTIONS

BRD and FP are two GT-rooted MAL schemes. Their mechanisms are discussed in the following, and their application to RAT selection is analyzed through literature examples.

1) BEST RESPONSE DYNAMICS (BRD)

BRD is a learning scheme tailored for complete information games. BRD players select the strategy maximizing their own utility at each game step, taking into account the strategies played by the others. At game step t, the updating rule for user n adopting *simultaneous* BRD (sBRD) with pure strategies is as follows:

$$\text{sBRD:} \quad a_{n,t} \in \{\mathsf{BR}_n\} | \boldsymbol{a}_{-n,t-1}, \tag{11}$$

that is, the player selects the strategy maximizing its utility given the strategies adopted by the others during the previous game step. sBRD applies to complete but imperfect information games where players cannot observe the strategies adopted by the others at present time t, while the strategy profile played in $t-1$ is of common knowledge. *Asynchronous* BRD (aBRD) is instead applicable to games where players do not act simultaneously (e.g., sequential games) with the following updating rule:

$$\text{aBRD:} \quad a_{n,t} \in \{\mathsf{BR}_n\} | [a_{1,t}, \ldots, a_{n+1,t-1}, \ldots, a_{N,t-1}]. \tag{12}$$

In this case, the perfect information assumption is verified, since each player observes the strategies of previous players also in the present game step t.

aBRD converges to PNEs when applied to EPGs [29] [31, Chapter 5], independently of the strategy profile at $t = 1$. No general convergence results exist for sBRD; as a matter of fact, [33] shows an unwanted, infinite strategy switching in a simple 2-player game caused by sBRD. However, the adoption of a stabilizing term can lead to PNEs [32], [158].

In RAT selection, aBRD is most often adopted, assuming a null probability of users performing selection at the same time. In particular, the work in [102], [103] shows convergence to the set of PNEs for CGs. Reference [102] also highlights that the number of iterations required by BRD to converge increases as a linear function of the number of users and available RATs (e.g., about 150 iterations for 200 users and 15 RATs). Moreover, convergence to pure ε-SPEs is shown when the game includes a previous step where RATs compete for frequency resources. In [108], BRD is used to solve the proposed CG, which has two available RATs and a unique PNE, and show convergence time increasing with the number of users (about 50 iterations with 250 users).

A BRD-inspired algorithm is proposed in [44] and shows convergence to PNEs in a game using the utility functions in (2a)(2b). On the one hand, the proposed procedure guarantees convergence when only one between the two functions is adopted, i.e., in case of homogeneous RATs; on the other hand, the use of a *hysteresis policy* enables convergence when

the two utility functions are used in parallel, ultimately solving scenarios with both WiFi and LTE RATs. Such a policy *stabilizes* the switching of users across classes of RATs, e.g., from WiFi to LTE and vice versa, avoiding the selection of a new RAT if the utility gain is lower than a predefined hysteresis value. The procedure is used as a benchmark in [46] for a selection game with a QoE-based utility function.

2) FICTITIOUS PLAY (FP)

A BRD player is able to observe the strategies selected by others *right before* its own move, e.g., at time $t-1$ for sBRD as reported in (11). FP expands the degree of observability to perfect recall, as defined in Section IV-B.

Since this assumption requires high information sharing and storage capability, FP has not been extensively used in wireless communication scenarios, and RAT selection makes no exception. However, it is discussed in this paper since, as clarified later, its updating rule provides the starting point of RL-based schemes (cf. Sections VI-B–VI-D).

Each FP player n is able to collect the history of pure strategies selected by the others, from a given time (e.g., $t=1$) to the instant at which it plays its own strategy. Then, at time t, the frequency with which players have selected their strategies up to $t-1$ is of common knowledge. For example, the selection frequency at time t for strategy a_n^i, denoted as $f_{n,t}(a_n^i)$ is:

$$f_{n,t}(a_n^i) = \frac{1}{t-1}\sum_{\tau=1}^{t-1} \mathbb{1}_{[a_{n,\tau}=a_n^i]}, \tag{13}$$

where $\mathbb{1}_{[\dots]}$ is the indicator function. The following recursive way to compute the selection frequency is also verified:

$$f_{n,t}(a_n^i) = f_{n,t-1}(a_n^i) + \alpha_{n,t}[\mathbb{1}_{[a_{n,t}=a_n^i]} - f_{n,t-1}(a_n^i)], \tag{14}$$

where $\alpha_{n,t} = \frac{1}{t}$ and regulates the *learning rate* of player n over time [32], [59]. Such a term has been generalized in other RL algorithms applied to RAT selection, as discussed later.

Once evaluated for all $a_n \in \mathcal{A}_n$, the selection frequencies are used to draw an empirical probability distribution $\pi_{n,t}^{\mathsf{emp}}(a_n^1,\dots,a_n^{|\mathcal{A}_n|}) = (f_{n,t}(a_n^1),\dots,f_{n,t}(a_n^{|\mathcal{A}_n|}))$, i.e., an estimate of the mixed strategy of player n at time t. Finally, given the empirical mixed strategy profile $(\pi_{n,t}^{\mathsf{emp}}, \boldsymbol{\pi}_{-n,t}^{\mathsf{emp}})$, the updating rule of player n is:

$$\text{FP:} \quad a_{n,t} \in \{\mathsf{BR}_n\} | \boldsymbol{\pi}_{-n,t}^{\mathsf{emp}}. \tag{15}$$

Hence, player n best responds to the mixed strategy profile estimated by observing the others, and selects the pure strategy that maximizes its utility. While a BRD player aims at maximizing the actual utility, the FP counterpart focuses on expected utility (as in (5), with π_n replaced by a_n).

The above description applies to simultaneous games. An asynchronous FP version can be applied instead to sequential games, similarly to BRD. A *smoothed* FP, also referred to as stochastic FP, has also been proposed, where players build a mixed strategy over time and apply it as an updating rule [29], [33]. Such a flavor of randomicity allows to avoid possible non-convergence issues of FP, that arise in games with cyclic behaviors. Hence, FP converges to PNEs if applied to PGs, since these are acyclic games [29] [31, Chapter 5]; it may not converge in other games, where, however, the following results are demonstrated [27, Chapter 5] [159]:

- PNEs are *attractors* to FP: if a PNE is played at game step t^*, then it is played for all $t > t^*$;
- The convergence to a pure strategy profile of a FP procedure guarantees the convergence to a PNE.

In the above formulation, each FP player derives the empirical strategy profile $(\pi_{n,t}^{\mathsf{emp}}, \boldsymbol{\pi}_{-n,t}^{\mathsf{emp}})$ assuming independent players. However, as discussed in Sections IV-D and IV-E, joint mixed strategy profiles allow to expand the set of game solutions, including CEs. A further FP version has thus been proposed and applied to stochastic games, where joint empirical probability distributions are estimated and used by users in their decision process [31, Chapter 5] [160].

When applied to RAT selection, FP convergence is challenged in [31, Chapter 5, Example 147], where an unwanted cyclic behavior in a 2-player and 2 RATs game is shown.

B. RL-BASED SOLUTIONS WITH MDP MODELING

From Section V-C, solving a MDP implies to deriving a policy that maximizes the agent return, i.e. (10) but with no dependence on other agents due to the single-agent assumption.[20] Different policies can be compared by means of the return for the agent, if applied. Considering an initial state s, from which a generic policy Pol starts to be applied, the so-called *state-value function*, denoted as $V^{\mathsf{Pol}}(s)$, can be evaluated and associated to Pol (for all $s \in \mathcal{S}$), as follows:

$$V^{\mathsf{Pol}}(s) := \mathbb{E}_{\mathsf{Pol}}\left[\sum_{t=0}^{+\infty} \lambda^t u_{t_0+t+1} \middle| s_{t_0} = s\right], \tag{16}$$

where $\mathbb{E}_{\mathsf{Pol}}[\cdot]$ indicates the expectation with respect to Pol, and $s_{t_0} = s$ means that the first state visited under Pol is s. While in a state, an either deterministic (pure) or stochastic (mixed) Pol drives the agent to select an *action*, that is, a strategy $a \in \mathcal{A}$. Then, for all possible (state, action) pairs, the so-called *action-value function* can be defined via corresponding Q-values denoted $Q(s,a)$. The Q-value is a measure of the overall expected return obtained when the agent is in state s and performs action a, and then continues to apply policy Pol. It is evaluated as follows:

$$Q^{\mathsf{Pol}}(s,a) := \mathbb{E}_{\mathsf{Pol}}\left[\sum_{t=0}^{+\infty} \lambda^t u_{t_0+t+1} \middle| s_{t_0} = s, a_{t_0} = a\right], \tag{17}$$

where $a_{t_0} = a$ indicates that a is the strategy applied by the agent at time t_0 while the environment is in s.

The Bellman equations allow to write both state-value and action-value functions in a recursive form [161]. For example, assuming a state/action-finite MDP, the Bellman equation for

[20] For this reason, the subscript n is not used in this section, and the lack of player-related subscripts implies the single-agent setup.

the state-value function is:

$$V^{\mathsf{Pol}}(s)=\sum_{a\in\mathcal{A}} p^{\mathsf{Pol}}(s,a)\sum_{s'\in\mathcal{S}} p(s'|s,a)[R(s'|s,a)+\lambda V^{\mathsf{Pol}}(s')], \tag{18}$$

where $p^{\mathsf{Pol}}(s,a)$ is the probability of adopting action a when the state is s, as defined by policy Pol, and $p(s'|s,a)$ is the state transition probability, defined as in Section V-C but in single-agent terms. Moreover, $R(s'|s,a)$ is the so-called *reward* function, representing the expected utility when a transition from s to s' occurs due to action a, that is:

$$R(s'|s,a)=\mathbb{E}[u_{t+1}|s_t=s,a_t=a,s_{t+1}=s']. \tag{19}$$

Given a policy, the system of Bellman equations for state-values (for all $s\in\mathcal{S}$) or action-values (for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$) has a unique solution [162]. Then, a policy is optimal if it maximizes state-value and action-value functions, so that it can be represented by the following values:

$$V^*(s) = \max_{\mathsf{Pol}} V^{\mathsf{Pol}}(s) \tag{20a}$$

$$Q^*(s,a) = \max_{\mathsf{Pol}} Q^{\mathsf{Pol}}(s,a) \tag{20b}$$

$$\Downarrow$$

$$V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a). \tag{21}$$

Writing Bellman equations for an optimal policy leads to the system of Bellman optimality equations [161], that has a unique solution, i.e., $V^*(s)$ or $Q^*(s,a)$ values for all states and actions. Hence, the system can be solved if the state transition probabilities are known, and one among all possible policies leading to optimal values can be identified. Based on the optimality equations, two aspects can be highlighted:

- The search for an optimal policy can be restricted to the set of deterministic and stationary policies, that map an action to each possible state, since there always exists at least one deterministic policy being optimal (under the infinite time-horizon discounted model) [163];
- The evaluation of $Q^*(s,a)$ is preferable while searching for the optimal policy, since the Bellman optimality equations for the action-value indicate that an optimal policy always chooses an action that maximizes the Q-value for a given state. Hence, an optimal policy is *greedy* with respect to Q-values.

By denoting a deterministic optimal policy as $\mathsf{Pol}^*(s)$, and observing a particular state s, the two above aspects result in the following equation:

$$\mathsf{Pol}^*(s)=\arg\max_{a\in\mathcal{A}} Q^*(s,a). \tag{22}$$

Solving Bellman equations to derive an optimal policy can be computationally expensive since the search space grows with state and action sets. Even under the assumption of knowing the environment dynamics, iterative methods are thus more suitable. As mentioned in Section V-C, dynamic programming algorithms such as Value and Policy Iteration are, among others, largely used, and converge to an optimal (deterministic) policy in case of finite MDPs. In case of unknown state transition distribution, RL algorithms, e.g., Q-Learning, provide a robust method to find an optimal (deterministic) policy; They iteratively solve approximated Bellman optimality equations, where the *experienced* state transitions are adopted in place of the unknown expected state transition distribution [164].

1) Q-LEARNING

Q-Learning is one of the most famous RL algorithm to solve finite MDPs under the assumption of unknown environment dynamics, i.e., unknown state transition probabilities and utility functions [39, Chapter 6] [164]. Q-Learning belongs to the temporal-difference (TD) methods, which combine dynamic programming with so-called Monte Carlo methods, the latter being another way of solving finite MDPs with no a priori knowledge [39, Chapter 5]. In particular, Q-Learning is a method that aims at learning optimal Q-values rather than state-values [164]. By denoting (s_t,a_t) the pair of state and action at time t, the Q-Learning rule to update the Q-value for this pair at time $t+1$ is:

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha_t\big[u_{t+1}+\lambda\max_{a\in\mathcal{A}} Q(s_{t+1},a)-Q(s_t,a_t)\big], \tag{23}$$

where s_{t+1} is the new state, u_{t+1} is the instantaneous utility, and $0<\alpha_t<1$ is the *learning rate*, that is assumed to be time-dependent. The rate regulates how the agent updates the estimate of $Q(s_t,a_t)$, considering both the old estimate and the most recent observation, that depends on the instantaneous utility and the *best* value of the newly encountered state $Q(s_{t+1},a)$ (maximized over all possible $a\in\mathcal{A}$). Small learning rates attribute more importance to the old estimates, while large ones give more credit to the newest observation. The max operator underlines the possibility for the agent to adjust its policy at each step. For this reason Q-Learning is also referred to as an off-policy method. An on-policy alternative scheme is the *SARSA* algorithm [165], [166].

Both Q-Learning and SARSA require the agent to track and update Q-values for all possible combination of states and actions. A popular approach is to embed the agent with a tabular memory, so that it can update over time its own Q-table. As also discussed in Section VI-B1.e, this approach hinders the scalability of the algorithms, and additional methods are needed in scenarios with large state and action spaces.

Q-Learning and SARSA converge to an optimal policy, as shown in [164], [167], [168] for Q-Learning, and [169] for SARSA. Two constraints are highlighted for Q-Learning:

- the learning rate has to *reasonably* decay over time;
- every state-action pair has to be visited infinitely often.

The first constraint implies:

$$\sum_{t=1}^{\infty}\alpha_t=\infty \tag{24a}$$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty. \tag{24b}$$

The above conditions balance the trade-off between the need for exploring, in the initial searching steps, vs. the goal of converging to an estimate, as time goes by. However, due to the need for fine tuning, and excessive convergence delays, the above conditions are used in theoretical derivations, and Q-Learning practical applications usually adopt a constant learning rate. This rate has to be small enough in order to avoid too large variations of Q-values in response to recent observations [39]. Note however that a continuous variation is preferable when dealing with non-stationary scenarios such as multi-agent situations.

The second constraint avoids the iterative procedure to get stuck into sub-optimal policies, that may seem optimal in the short-term but are not. Hence, it allows for alternating exploration and exploitation steps. In practice, when an action must be selected, the agent cannot *always* be greedy with respect to the evaluated Q-values (full exploitation), e.g., by applying (22) at each t; it must adopt instead a procedure that allows to keep exploring other strategies over time [36], [38]. Therefore, ϵ-*greedy* policies are widely adopted in combination with Q-Learning, where a parameter $0 \leq \epsilon < 1$ defines, at each state (time), the probability ϵ of the agent exploring a randomly selected action while being greedy toward the best (up to that moment) action with probability $1 - \epsilon$. In stationary conditions the exploration is particularly needed at the beginning of the process; hence, ϵ-greedy procedures with ϵ vanishing over time have also been proposed. Another option is to use the so-called Boltzmann procedure (BP), that selects action a' in state s with probability $p^{\mathsf{BP}}(s, a')$ defined as follows:

$$p^{\mathsf{BP}}(s, a') := \frac{e^{Q(s,a')/\mathsf{T}}}{\sum_{a \in \mathcal{A}} e^{Q(s,a)/\mathsf{T}}}, \tag{25}$$

where the higher the so-called *temperature* $\mathsf{T} > 0$, the higher the randomicity of the exploration. A soft greedy exploitation is preserved with BP, since actions with higher Q-values have greater chances of being selected for all $\mathsf{T} \in (0, \infty)$. A detailed analysis of exploration strategies for MDP policies can be found in [170].

The above discussion identifies Q-Learning as a powerful MDP solver, and also shows its applicability to multi-agent scenarios. Nothing prevents agents to apply Q-Learning independently of one another; this directly maps to stochastic games with imperfect and incomplete information, where players may be unaware of being part of a game. Using Q-Learning as a stochastic game solver is thus possible. However, it leads to uncertain solutions, since convergence to an optimal policy profile is not guaranteed due to non-stationarity. The simplicity of this approach has triggered its widespread use, also in RAT selection. A set of examples of application of Q-Learning to RAT selection is reported and discussed below. The literature is also summarized in Table 4.

a: Q-LEARNING AS MAIN ALGORITHM OR BENCHMARK

One of the first applications of Q-Learning to user-centric RAT selection is found in [96]. The model intersects evolutionary and stochastic games, since the adopted utility function is a user-independent version of the PF model in (2b), with a further variable connection price (Section III-B2). The possibility to converge to the evolutionary equilibrium (where all users have same utility) is explored via either a network-assisted RD-based strategy or a fully distributed Q-Learning mechanism. For the latter, a simplified version is actually used, where a Q-value is associated with each possible pure strategy (a candidate RAT), with no explicit state definition: users only observe their utility obtained when they connect to a RAT. ϵ-greedy is used to balance exploration and exploitation. The analysis in [96] is carried out in a scenario with a fixed number of users (variable in some specific analyses) and RATs, i.e., a WiFi AP, a code division multiple access (CDMA)-based macrocell, and an IEEE 802.16 (WiMAX) base station (BS). It is shown that Q-Learning converges to the evolutionary equilibrium and convergence is obtained with a higher number of iterations compared to the network-assisted solution, that however exploits broadcast information about users and utility. A more detailed analysis shows that the evolutionary equilibrium is reached for a specific ϵ value.

A similar study is carried out in [109]; in a cellular scenario including a macrocell and two, open- vs. closed-access femtocells,[21] fully greedy Q-Learning ($\epsilon = 0$) achieves a sub-optimal solution compared to RD.

The scheme in [96] has been widely used as a benchmark for other methods proposed over the years. In [47], it is compared against the so-called local improvement algorithm (LIA) and its enhanced version (E-LIA) in a selection scenario where users have different utility. Utility is a function of the user physical rate, adjusted with some coefficients depending on required traffic (brittle, partially elastic, and elastic) [81]. LIA introduces cooperation between pairs of access nodes, each referred to as coupled network pair (CNP), that share users by following an optimization strategy. Exploiting the spatial distribution of access nodes and assuming a priori knowledge of user required traffic, LIA decomposes a global welfare optimization problem into low-complexity sub-problems, where each CNP cooperatively re-associates users exploiting the knowledge on their traffic. E-LIA speeds up LIA convergence using the spatial independence among CNPs. Simulations in [47] are performed for the topologies in Section III-C and show that LIA and E-LIA outperform Q-Learning, with a welfare improvement depending on the weights adopted to characterize the traffic classes. Q-Learning shows increasing issues in heavily-loaded scenarios, i.e., about 100 users for the Corridor scenarios and 60 for Overlapping and Nest scenarios.

[21] Open access cells allow all users to access, while closed access ones are dedicated to pre-registered users. Moreover, all users are allowed to access to hybrid access cells, but registered users have higher priority [182].

TABLE 4. Examples of application of *Q*-Learning to RAT selection.

Reference	Distributed/ Centralized/ Hybrid	Main Algorithm/ Benchmark	Q-Learning Settings	Scenario Settings and Topology
[96]	Distributed	Main Algorithm	$\epsilon = 0.1$ $\alpha = 0.1$ $\lambda = 0.2$	$\|\mathcal{N}\| = 50$ (it varies for some analyses) States not defined $\|\mathcal{A}_{\text{tot}}\| = 3$ Nest: 1 Macrocell, 1 WMAN, 1 WLAN
[171]	Distributed	Main Algorithm	ϵ-adaptive $\alpha = 0.5$ $\lambda = 0.9$	$\|\mathcal{N}\| = 10$ State not defined $\|\mathcal{A}_{\text{tot}}\| = 31$ Nest: 1 Macrocell, 30 Small cells
[172]	Distributed	Main Algorithm	$\epsilon = 0.1$ $\alpha = 0.5$ $\lambda = 0.5$	$\|\mathcal{N}\| = 75$ State is Top-2 received powers $\|\mathcal{A}_{\text{tot}}\| = 3$ Nest: 1 Macrocell, 2 Small cells
[109]	Distributed	Main Algorithm	$\epsilon = 0$ $\alpha = 0.2$ $\lambda = 0.2$	$\|\mathcal{N}\| = 110$ State not defined $\|\mathcal{A}_{\text{tot}}\| = 3$ Nest: 1 Macrocell, 2 Small cells
[173]	Distributed	Main Algorithm	$\epsilon = 0.35, 0.1$ $\alpha = 0.3$ $\lambda = 0.7$	$\|\mathcal{N}\| = 3$ State is combination of user/network indicators $\|\mathcal{A}_{\text{tot}}\| = 2$ Overlapping: 2 Macrocells
[174]	Distributed	Main Algorithm	$\epsilon = -$ $\alpha = 0.5$ $\lambda = 0.9$	$\|\mathcal{N}\| = 1$ State is position and available networks $\|\mathcal{A}_{\text{tot}}\| = [15 : 60]$ Nest: 1 Macrocell, $\|\mathcal{A}_{\text{tot}}\| - 1$ WLANs
[175] [48]	Distributed	Main Algorithm	$\epsilon = 0.4, 0.6$ or $0.1, 0.3$ $\alpha = 0.3$ $\lambda = 0.3$	$\|\mathcal{N}\| = [1 : 6]$ State is traffic type, available networks, time $\|\mathcal{A}_{\text{tot}}\| = 3$ or 4 Overlapping: 1 Macrocell, 1 or 2 WLANs, 1 VLC
[176]	Distributed	Main Algorithm	$\epsilon = 0.1$ $\alpha = 0.9$ $\lambda = 0.9$	$\|\mathcal{N}\| = 15$ State is available networks, packet success ratio, load $\|\mathcal{A}_{\text{tot}}\| = 5$ Nest: 2 Macrocells, 3 WLANs
[47]	Distributed	Benchmark	Not reported (Ref. [96])	$\|\mathcal{N}\| = 110$ State not defined $\|\mathcal{A}_{\text{tot}}\| = [3, 4, 6]$ Nest: 3 Macrocells and WLANs topologies
[46]	Hybrid	Benchmark	Not reported (Ref. [96])	$\|\mathcal{N}\| = [54, 68, 98]$ State not defined $\|\mathcal{A}_{\text{tot}}\| = 4$ Nest: 2 Macrocells, 2 Small cells
[108]	Distributed	Benchmark	Not reported (Ref. [96])	$\|\mathcal{N}\| = [15 : 40]$ State is # of users connected to each network $\|\mathcal{A}_{\text{tot}}\| = 2$ Nest: 1 Macrocell, 1 Small cell
[177]	Distributed	Benchmark	Not reported (Ref. [171])	$\|\mathcal{N}\| = [5 : 30]$ State not defined $\|\mathcal{A}_{\text{tot}}\| = 2$ Nest: 1 Macrocell, 1 Small cell
[178]	Distributed	Benchmark	Not reported (Ref. [96])	$\|\mathcal{N}\| = [500 : 800]$ State not defined $\|\mathcal{A}_{\text{tot}}\| = 2$ Nest: 1 Macrocell, 1 WLAN
[179]	Centralized	Main Algorithm	$\epsilon = -$ $\alpha = 0.1$ $\lambda = 0.95$	$\|\mathcal{N}\|$ variable State is interference level and busyness ratio $\|\mathcal{A}_{\text{tot}}\| = 3$ Nest: 1 Macrocell, 2 WLANs
[180]	Hybrid	Main Algorithm	ϵ variable α variable λ variable	$\|\mathcal{N}\|$ variable State is # users in a given area and QoS requirement $\|\mathcal{A}_{\text{tot}}\| = 3$ Overlapping: 1 Macrocell, 2 WMAN
[181]	Centralized	Main Algorithm	$\epsilon = -$ (also BP and VDBE) $\alpha = 0.5$ $\lambda = 0.1$	$\|\mathcal{N}\| = [10^5 : 10^6]$ State is the selected macrocell $\|\mathcal{A}_{\text{tot}}\| = 209$ Overlapping: all Macrocells (real data or Poisson Point Process)

In [46], Q-Learning is analyzed against a hybrid scheme adopting collaboration and information exchange between users and a dedicated cloud entity. The goal is to achieve a PNE, which is proven to exist from a QoE perspective. Simulation results show competitive performance of Q-Learning, but it is claimed that it may take several hundreds of iterations to converge even in scenarios with a low number of users.

The Q-Learning scheme in [96] is also adopted in [178] to benchmark a learning scheme referred to as ALA. The scenario includes WiFi, WiMAX, and OFDMA LTE-like RATs. The utility is defined as a weighted logarithmic function of the ratio between experienced and requested throughput, with the latter depending on the service (voice vs. data). The throughputs equate when the sum of the demands of users choosing the same RAT is lower than RAT capacity. Otherwise, the experienced throughput is equal to capacity multiplied by the ratio between the user demand and the sum of the demands of the other users. The capacity also changes over time in a gradual (sinusoidal) vs. abrupt fashion. Overall, the proposed utility formulation emphasizes the concept of user *satisfaction*, that is dependent on the requested service. Simulation results include several performance indicators, such as user utility, switching rate, bandwidth utilization, and convergence time, adopting various number of users (from 500 to 800) in both scenarios of gradually- vs. abruptly-changing capacity. It is shown that Q-Learning needs more iterations than ALA to converge to a load-balanced configuration (that is a PO and SO PNE for the proposed game). However, ALA embeds a quite complicated forecasting method requiring each user to store the historical series of loads over the selected RATs (obtained via network assistance).

A model similar to [96] is proposed in [171] and adopted in a scenario including one macrocell and several femtocells. An *adaptive-*ϵ exploration strategy is proposed and used in the single-agent scenario, which forces the agent to explore more when the knowledge about the environment is uncertain. BP is adopted instead in the multi-agent setup. The reward is the Shannon's capacity when a RAT is selected, with a bandwidth equally split across users. Convergence to stable configurations is not analyzed, and results are given in terms of instantaneous vs. average user throughput. The user adopts either Q-Learning or a scheme where it selects the cell initially leading to highest capacity.

In [108], BRD is compared against a Q-Learning solution where ϵ is randomly picked, at each time step. The selection of a RAT is the action, while the number of users connected to each RAT is the state. The problem of sharing the state information among users is not discussed. The utility function is similar to [96] with a further fixed connection price. This choice enables a further comparison against RD. Simulation results show that Q-Learning converges to the unique game PNE; it takes more iterations to converge with respect to BRD and RD. The number of steps before convergence increases with the number of users, with more than 3×10^4 steps to converge in a scenario with 40 users and 2 RATs.

The methods proposed in [171] and [108] are used to benchmark the network-assisted solution proposed in [177], where users periodically perform RAT selection by using *association probabilities* evaluated by a central unit. Users connect to a RAT for a given amount of time and then perform a new selection. In the meantime, the central unit computes the association probabilities and provides them to users, so that they can apply it in the next period.

A Q-Learning selection approach with ϵ-greedy exploration is also proposed in [172] in a macro vs. picocells scenario. Here, the so-called cell range expansion (CRE) technique is also considered. Aiming at load balancing, CRE allows to bias the value of the power received from the picocells, thus influencing users to use them [183], [184]. The state is user-specific since it includes, for each user, the two power levels received from the macrocell and the best picocell at a given time. Moreover, the selection of a cell defines the action, and a cost is adopted to update the Q-values.[22] The cost is defined as the number of users unable to achieve the connection, referred to as UE outage. The cost values are exchanged across cells over their backhaul connection and then reported to users. Such a definition indirectly triggers cooperation between users in a non-cooperative stochastic game. Finally, in order to deal with the size of the Q-table, the states are quantized between an upper and a lower bound, and a new state is added only if it does not fit the table of the previous time step. The convergence is not analyzed and the throughput experienced after a fixed number of trials (5×10^5) is considered as the performance indicator. The proposed scheme outperforms plain Q-Learning but also a scheme where users connect to cells using common and predefined CRE values. An optimal but unpractical approach is also used as upper bound, where the CRE values minimizing the number of UE outages are found via exhaustive search.

Q-Learning is combined with offline unsupervised ML (clustering) and online supervised ML (classification) in [173]. X-means is used to cluster a given amount of training data, that identify possible states of the candidate RATs in terms of indicators, such as, load and DL SINR. Then, the current user readings are associated with a cluster via k-nearest neighbors (kNN). Once the appropriate cluster is selected, Q-Learning is used to obtain the best action, that is, the best RAT to use. The reward is a function of experienced throughput and load, the latter being shared by RATs. In a scenario with two LTE cells and three random walking users, Q-Learning outperforms a selection scheme based on the highest SINR, as well as a random selection procedure. It is also shown that the random procedure, usually adopted for showing performance lower bounds, surprisingly outperforms the SINR-based scheme.

Reference [174] adopts ϵ-greedy Q-Learning for cellular to WiFi offloading. A mobile user learns how to select between a macrocell and randomly deployed WiFi APs, using its connection history and current network states. The goal is

[22] Hence, users select minimum Q-values during exploitation.

to decrease the connection time toward the macrocell while maintaining a minimum QoS. The user moves through specific positions where the selection is performed. The state at time t is defined as a tuple including the position and available WiFi APs, obtained via 3GPP ANDSF. The macrocell utility is the SINR experienced upon connection, while a more complicated utility function is defined for the WiFi APs. It considers SINR, handover delay, load, and an incentive factor. The load is obtained via the channel busy fraction parameter available in the IEEE 802.11k standard. Then, it is weighted depending on the interest of the cellular operator, i.e., the operator may want to increase offloading toward WiFi when the macrocell is congested. The incentive factor is defined as a logarithmic function of the inverse of the distance between user and macrocell. Q-Learning (unknown ϵ, $\alpha = 0.5$, and $\lambda = 0.9$) is evaluated against a selection scheme based on received signal strength (RSS) and load. Results show the possibility to decrease the cellular connection time by appropriately tuning the weight associated to the load. Then, the iterations to converge to the optimal policy increase with the number of APs (about 50 iterations with 60 APs). Note that in this investigation the implications related to a multi-agent setup are not considered.

b: Q-LEARNING ENHANCEMENTS

A Q-Learning enhancement via *knowledge transfer* is proposed in [48], [175], aiming at reducing Q-Learning action set by using past experiences. This enhancement is tested for both DL and UL RAT selection in a hybrid visible light communication (VLC), LTE, and WiFi access system. PF and TF throughput models are adopted for LTE and WiFi, while a specific function is used for VLC. Then, DL and UL throughput indicators are merged in different ways in order to define the user utility in UL-dominant, DL-dominant, and UL/DL-symmetric traffic scenarios. Q-Learning is enhanced by considering that: a) not all RATs are suitable for all traffic types, and b) network load has space-time patterns. The first observation allows to decrease the size of the action set (composed by the candidate RATs) according to traffic type, e.g., VLC is not suitable for UL data. The second observation highlights that the load at a given time and location is about the same across different weekdays. Hence, a user can reuse past experience and Q-values previously evaluated, instead of initiating the RAT selection with null Q-values every time. The user can also differently adjust its exploration probability if previous knowledge is exploited, i.e., there is no need for large exploration probabilities at the beginning of the process. Simulation results show that the improved Q-Learning converges more rapidly than its legacy counterpart.

Another Q-Learning enhancement is given in [176]. The proposed model-driven framework intersects ML and GT and consists of feature learning, game modeling, and strategy learning steps; overall, it is referred to as random forest enhanced Q-Learning with game (RFEQG). For each user, the action set comprises the available RATs while the state is a tuple of three elements: the list of RATs with corresponding load and packet success ratio (PSR). The load is obtained via signaling mechanisms exploiting ANDSF or IEEE 802.11u (aka WiFi Alliance Hotspot 2.0) functionalities [185]. PSR is instead derived at user side via the random forest supervised learning algorithm. The latter is used to model the relationship between PSR and some features experienced by the users, such as RSRP [dBm], bit error rate (BER) [%], and SINR [dB]. The algorithm is tested in a mixed LTE/WiFi scenario, adopting the utility functions in (2). RFEQG users decide the action via ϵ-greedy Q-Learning. However, before executing the strategy, they estimate the possible improvement and actuate the decision only in case of a significant estimated gain. This procedure is inspired by the hysteresis policy adopted in [43], [44] to stabilize BRD (Section VI-A1). When applied to Q-Learning, it rectifies excessive RAT switchings, thus speeding up convergence towards a PNE. In a scenario with 4 RATs (2 LTE macrocells and 2 WiFi APs) and 15 users in a 80×80 m^2 area, results show convergence and utility improvements of RFEQG (and intermediate algorithms from Q-Learning to RFEQG) with respect to a traditional Q-Learning ($\epsilon = 0.1$). However, it is unclear how the state is defined for Q-Learning, since PSR modeling via random forest seems to apply to RFEQG only.

c: Q-LEARNING-BASED HYBRID AND CENTRALIZED SELECTION

The work in [179]–[181] provides further examples of applying Q-Learning to RAT selection. However, Q-Learning is used here for hybrid and centralized selection. A short description of these works is given for completeness.

In [179], Q-Learning based network selection (QBNS) is proposed and tested in a wideband CDMA (WCDMA)/WLAN scenario. QBNS takes into account network capacity and QoS requirements of each user. When a new user enters the system, a central entity implements QBNS, where the state is a combination of WCDMA interference level and WLAN busyness ratio, and the action is the selection of a RAT type. The reward is a linear combination of network and user utility, with the latter considering the required service and the QoS in terms of data rate, delay, and BER. Simulation results in a scenario with one WCDMA, two WLANs, and a variable user arrival rate, show that QBNS outperforms an algorithm based on a semi-Markov decision process (SMDP) in terms of average number of users admitted in the system, voice call blocking probability, and discounted reward. Adopting $\alpha = 0.1$ and $\lambda = 0.95$, QBNS converges in about 100 iterations.

A SMDP model is also adopted in [180] and solved by Policy Iteration vs. ϵ-decaying Q-Learning in complete vs. incomplete information cases. In SMDPs, actions have a continuous time duration with respect to discrete and/or fixed duration in MDPs. Reference [180] proposes SMDP to model a network-assisted scheme, where networks learn how to provide appropriate information, in terms of QoS parameters and costs to be paid per amount of traffic, aiming at encouraging/discouraging users to connect. The information

is shared via logical signaling channels defined in IEEE 1900.4 [186]. The goal is to meet operator objectives, while users maximize their utility. Users explicitly take the final decisions, and thus the proposed scheme is a hybrid RAT selection approach. Compared to Policy Iteration, and a simpler policy called *staircase*, where the shared parameters are not dynamically derived but optimally pre-configured, Q-Learning shows lower average user throughput and higher blocking probability.

A MDP model for centralized RAT selection in cellular massive machine type communication (mMTC) is proposed in [181]. In this case, a dedicated central entity selects a cell for each MTC device. The state is the cell selected in the previous time step. The reward is a linear combination of the load of the selected cell and the device data rate, which is zero in case of connection blocking. Q-Learning is applied with three exploration schemes: ϵ-greedy, BP (see (25)), and value-difference based exploration (VDBE), that allows to define state- and time-dependent exploration probabilities [187]. Simulations adopt cell topologies based on real data[23] and Poisson point process (PPP). In both cases, VDBE outperforms the other exploration strategies in terms of blocking probability and average transmission rate. Compared against the highest-RSS selection scheme, Q-Learning shows a lower blocking probability and achieves load balancing across cells; however, it struggles in terms of average rate.

d: Q-LEARNING GOES MULTI-AGENT – NashQ

Q-Learning originally targets single-agent scenarios and its optimality and convergence are not demonstrated in a multi-agent setup. It basically neglects the multi-agent nature, minimizing in this way the need for information exchange and observability among agents. However, several Q-Learning extensions tailored to multi-agent systems exist in the literature along with other MARL algorithms [36], and can also be applied to RAT selection games. An example is found in [188], where the so-called *NashQ*-Learning algorithm [50], [189], referred to as *NashQ* below, is used to solve RAT selection. Each *NashQ* agent observes actions and utility of the others (i.e., the game is of perfect information), making it possible to create and update a Q-table dedicated to each opponent. This enables, at each game step, the computation of a MNE for the game matrix induced by the set of Q-tables in a given state. The Q-values in each Q-table are then updated as follows:

$$Q_n(s_t, \boldsymbol{a}_t) \leftarrow Q_n(s_t, \boldsymbol{a}_t) + \alpha_{n,t}\left[u_{n,t+1} + \lambda Q_{\mathsf{Nash}}(s_{t+1})\right], \tag{26}$$

for all $n \in \mathcal{N}$, where $Q_{\mathsf{Nash}}(s_{t+1})$ indicates the Q-value in the next state s_{t+1}, when a MNE strategy is adopted in t. $Q_{\mathsf{Nash}}(s_{t+1})$ can be evaluated assuming the observation of the Q-values of each player for each state, and is formally defined as follows:

$$Q_{\mathsf{Nash}}(s_{t+1}) := Q_n(s_{t+1}, \boldsymbol{\pi}^{\mathsf{NE}}). \tag{27}$$

The convergence of *NashQ* to NEs is discussed in [50], [189] and further analyzed in [27, Chapter 5].

In [188], RAT selection is obtained via *NashQ* in a multi-agent setup, but a single-agent approach based on Q-Learning is also discussed. In the second case, all users within an area with three different RATs (4G, 5G, and IEEE 802.16m, known as WiMAX Release 2) are considered as a single learning agent, and the services they require form a queue. Within each learning episode, the services are selected from the queue, one at a time, and associated to a RAT via ϵ-greedy Q-Learning. Then, a new episode initiates with a new service order. The states are defined as the tuples of network available capacity ratios (ACRs), defined as the ratios between currently available and total capacities. In order to narrow down the state space to sixty-four elements, the ACRs are quantized over four different levels. The utility is defined as the product between ACR and user preference in choosing a specific RAT. The latter depends on QoS requirements and RAT attributes, with a mapping obtained via AHP. In the multi-agent scenario, users are treated separately and *NashQ* is used instead of Q-Learning. Simulation results adopting $\epsilon = 0.5$ (surprisingly high and not decreasing over time), $\alpha = 0.1$, and $\lambda = 0.8$, show both Q-Learning and *NashQ* to better balance users over the available RATs with respect to a) a fair yet inefficient random selection, and b) a selection biased toward 5G.

The work in [190] integrates *NashQ* in the so-called smart aggregated RAT access (SARA), that aims at maximizing the long-term throughput of users in a cellular/WiFi hybrid system. Considering a scenario with users having different QoS requirements, SARA employs *NashQ* to solve a game of RAT plus channel selection across the subflows of a single user. Hence, the game players are the user's subflows, and *NashQ* is used to provide a set of feasible RAT/channel selection strategies for each of them. Then, a Monte Carlo tree search (MCTS) algorithm is sequentially adopted across users to perform the final selection for all of them. Numerical results in a scenario with less than 8 users and 2 candidate RATs show that SARA significantly improves network throughput compared to traditional WiFi offloading schemes, while guaranteeing traffic QoS requirements.

More recently, *NashQ* has been adopted in [92] to solve RAT selection in a network-centric approach that also embeds a user-centric perspective. The proposed framework is analyzed in a heterogeneous scenario with 5G, LTE-A, and WiFi available RATs, and users requesting three different service types (smart health, virtual/augmented reality, and industrial machinery). Given a service type, AHP and GRA are used to evaluate the best RAT to connect. This can be done at the user side by a) considering service requirements in terms of network attributes (throughput, energy efficiency, delay, jitter, packet loss rate, and price), b) evaluating the weights for each attribute via AHP, and c) using such weights in GRA for

[23] "Open cell ID", http://opencellid.org/, Accessed on: May 2021.

deriving a weighted grey correlation coefficient for each candidate RAT. The coefficient considers the RAT characteristics in terms of the above attributes, and indicates the preference of selecting a RAT for a specific service type. Each RAT is a *NashQ* agent and, at each step, decides on the service types (and corresponding users) to serve. Upon decision, RATs evaluate their utility as the difference between the total throughput of served users and blocking costs. Simulation results highlight the convergence to stable selection policies of the proposed framework. In comparison with the approach in [176], the proposed scheme shows similar throughput and energy efficiency performance, but lower user blocking probability and delay.

e: Q-LEARNING GOES DEEP – DEEP Q-LEARNING (DQL)

A well-known RL problem is its scalability. In Q-Learning, the main issue is the need for maintaining and updating over time the Q-table, the dimension of which grows with state and action sets. *NashQ* adds further complexity, since each agent must operate on N different Q-tables and evaluate a MNE strategy at each time step. As clear from Table 4, this challenge has been ignored when applied to RAT selection, by either reusing the *stateless* approach proposed in [96] or by limiting the size of both state and action sets. Nowadays, such solutions are questionable, particularly in 5G and beyond scenarios of massive connectivity and network/user/service heterogeneity, including enhanced mobile broadband (eMBB), ultra reliable low latency communication (URLLC), mMTC, and more general Internet of Things (IoT) paradigms. In these cases, RL finds a good ally in deep learning [191].

A deep learning agent extrapolates high-level models from data, without human intervention. The basic deep learning component is a neural network (NN) with multiple hidden layers, i.e., a deep NN (DNN). Starting from DNN, several architectures have been proposed, including feedforward neural networks (FNNs), recurrent neural networks (RNNs), convolutional neural networks (CNNs), generative adversarial networks (GANs), auto-encoders (AEs), etc. The description of these architectures is out of the scope of this work. The reader can refer to [191] for an introduction; comprehensive surveys on deep learning and its application to networking can be found in [192].

DRL refers to the set of methods that approximate value and action functions through DNNs [54], [55]. A DRL agent is a DNN that continuously interacts with the environment and receives feedback from it. As for RL, the agent takes actions that trigger the environment into new states. The training goal of the DNN is to optimize its parameters so that the agent can reliably select the actions leading to highest returns. DRL suits problems with a huge number of possible states, thus broadening the application of RL to high-dimensional scenarios. Inspired by the achievements in other fields [193], [194], DRL is under investigation for application to networking and communication problems, ultimately opening novel research perspectives. A comprehensive survey is given in [195] that also includes an exhaustive description of deep Q-Learning (DQL) schemes, that are the main DRL actors nowadays along with *deep policy gradient* methods [193], [196]. Hence, the reader can refer to [195] for more details on DQL. The main features of DQL are shortly summarized below for the purpose of introducing its application to RAT selection. The focus on DQL is due to its wider use for solving user-centric RAT selection, compared to deep policy gradient methods. Reference [24] provides a recent example of the use of the deep deterministic policy gradient (DDPG) algorithm, rather than DQL, for solving RAT selection in the context of heterogeneous health systems embedded with 3G, LTE, and WiFi RATs. Note that DDPG is used in [24] since it can deal with the continuous action space defined in the model, i.e., the percentage of data to transmit on each RAT, and the compression ratio.

As for Q-Learning, in DQL the action is usually selected by adopting an ϵ-greedy policy; however, DQL replaces the Q-table with a so-called deep Q-Network (DQN), aiming at approximating Q-values for each possible (state, action) pair [193]. DQN embeds a DNN, i.e., the *Estimation network*, that is characterized by a continuously updated vector of weights $\boldsymbol{\theta}$, and provides Q-value approximations, so that $Q(s, a; \boldsymbol{\theta}) \approx Q(s, a)$. In order to stabilize DNN learning and avoid local minima, a parallel *Target network* with vector $\boldsymbol{\theta}^{\mathsf{T}}$ is updated every N^{step} steps as a copy of the estimation network, so that $\boldsymbol{\theta}^{\mathsf{T}} \leftarrow \boldsymbol{\theta}$. In this setup, the optimization goal is to minimize the following loss function:

$$\mathcal{L}_t^{\mathsf{DQN}}(\boldsymbol{\theta}_t) = \mathbb{E}\Big[\big(u_{t+1} + \lambda \max_{a \in \mathcal{A}} Q(s_{t+1}, a; \boldsymbol{\theta}_t^{\mathsf{T}}) - Q(s_t, a_t; \boldsymbol{\theta}_t)\big)^2\Big], \tag{28}$$

where the expectation is with respect to the observed $(s_t, a_t, u_{t+1}, s_{t+1})$ tuple at time $t \rightarrow t+1$. The term between parentheses is also often referred to as *TD error*.

Note that the tuples experienced during learning are stored in a so-called *Experience Replay* memory $\mathcal{D}$. Then, during the learning process, the estimation network is trained using a random sample of tuples from $\mathcal{D}$ (often referred to as mini-batch), instead of only using the current experienced tuples. Figure 9 depicts, on a high level, the functioning on both Q-Learning (Figure 9a) and DQL (Figure 9b) for a direct comparison between the original algorithm and its deep evolution. Starting from the above general architecture, several variants have been proposed.

DQN With Prioritized Experience Replay (PER) [197]: Experience Replay minimizes correlation between the tuples used to train the DQN, by randomly using past tuples from $\mathcal{D}$ instead of current ones. PER allows to recall particular tuples, i.e., those for which the estimation did not perform well, since these are the tuples there is still something to learn about.

Double Deep Q-Network (DDQN) [198]: Rooted in double Q-Learning [39, Chapter 6] [199], DDQN separates action selection and action evaluation, to reduce possible over-estimation of Q-values during the training process.

Dueling DQN [200]: Q-values express how good it is to take a certain action in a given state. Hence, they can be

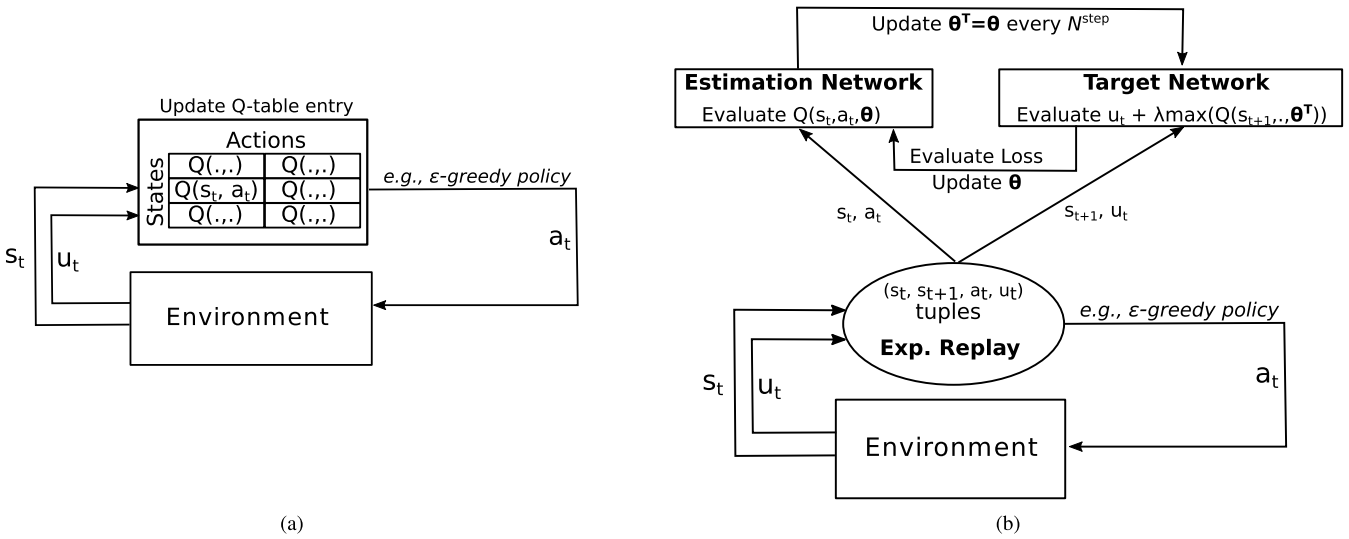

FIGURE 9. General structure and operations for tabular *Q*-Learning (a) and DQL (b). In (a), the updating rule follows (23); in (b), the loss function follows (28).

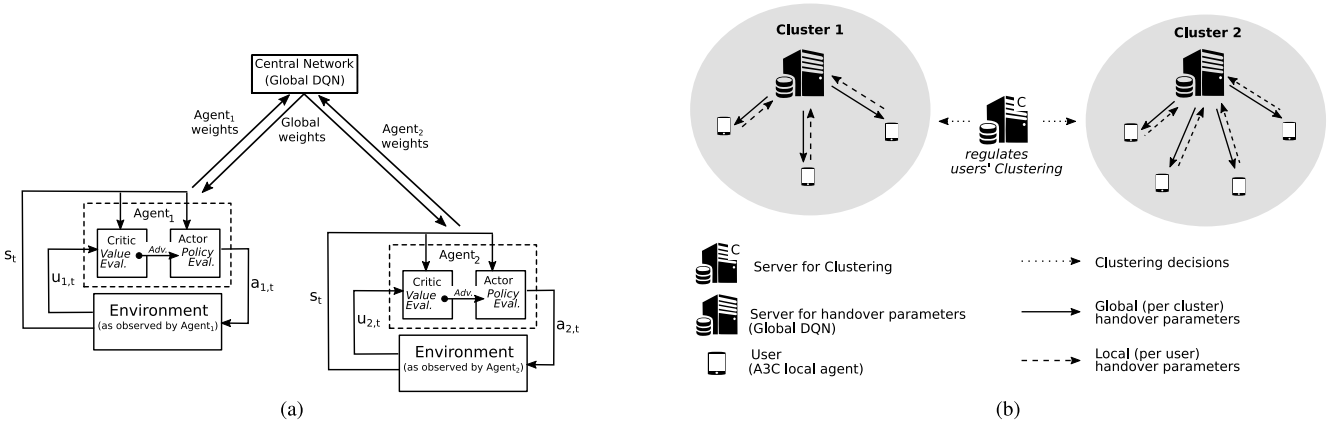

FIGURE 10. A3C general model (a) and application to RAT selection (handover) as proposed in [97] (b).

decomposed into two terms, so that $Q(s, a) = V(s) + A(a)$, where $V(s)$ estimates the importance of being in a particular state s and $A(a)$ estimates the gain in selecting action a compared to others. Moreover, *Q*-Learning and D*Q*N struggle in *redundant* situations [56], where two or more actions can be selected without getting any negative result. In D*Q*N, such situations can be addressed via the *dueling network* mechanism, where the D*Q*N estimation network is split into two sub-networks: the first, parameterized by $\boldsymbol{\theta}_1$, estimates $V(s; \boldsymbol{\theta}_1)$ while the second, with weights in $\boldsymbol{\theta}_2$, focuses on $A(s, a; \boldsymbol{\theta}_2)$. The two networks and corresponding estimates are then aggregated to approximate the *Q*-values.

Asynchronous Advantage Actor-Critic (A3C) [201]: A3C is an example of asynchronous RL, that tackles the same problems addressed by Experience Replay, by proposing the use of multiple (virtual or physical) agents in parallel, interacting with their own version of the environment (*asynchronous*). Moreover, similarly to the dueling scheme, the network of each agent is decomposed in two sub-networks: the *actor* estimates the *Q*-values, i.e., the policy, while the *critic* evaluates the value function. The critic indicates to the actor how good the selected action was, by evaluating and sharing the difference between the value function and the discounted return over the last T steps (*advantage*). Hence, the Actor-Critic method combines the benefits of value-iteration methods, such as *Q*-Learning, and policy-iteration methods, since the advantage mechanism originates from the latter [196]. Finally, the parameters computed independently by each agent are shared in a fully asynchronous manner[24] to a global DNN, that performs the global optimization and communicates the results to the agents [57]. A3C functioning is depicted in Figure 10a.

The above D*Q*N variants have been used in early works exploiting D*Q*L in RAT selection. More details on the same

[24] A synchronous version of A3C also exists and is referred to as A2C.

and other variants can be found in [56], [57], [195]. In the latter works, the multi-agent perspective is also introduced, since it is becoming increasingly popular and is also key for next-future analysis of RAT selection.

A mechanism for cellular handover is proposed in [97] aiming at minimizing energy-inefficient handovers in a dense deployment. A two-layer framework is proposed: the first layer is regulated by a centralized controller that groups users into clusters via unsupervised learning. Users in a same area and with similar mobility patterns form a group. In the second layer, users within each cluster independently perform handovers across the available cells. Each user resembles an A3C agent and learns its version of the optimal handover policy for the entire cluster. It thus shares the learned parameters in a cluster-dedicated server, that ultimately derives the parameters for the global policy. When a new user enters the cluster, it fetches from the server the most recent copy of the handover policy and then starts its own A3C cycle. Hence, it does not start the policy discovery from scratch. The framework is depicted in Figure 10b. The model is completed by defining an action corresponding to the selection of a cell, and as a state the RSRQ from the available cells. The reward is defined as a weighted linear combination of user data rate and handover energy consumption. The framework is tested in a scenario with three 16×16 m^2 areas, each covered by six cells randomly deployed. Four walking users per area are assumed. Results show that A3C achieves higher throughput and lower handover rate than the algorithm proposed in [202], based on a MAB model and solved via the upper confidence bound algorithm (UCB1) (cf. Section VI-C).

In [203], cellular to WiFi offloading is solved via DQN, aiming at minimizing user costs and energy consumption. Users pay a penalty if their data transmissions do not end before a given deadline. The state includes the user location and the remaining size for all active data flows. The user decides to transmit over WiFi or cellular RATs, or remain inactive (idle) in a given time step. It also decides how to allocate the flows on the channels available per RAT. Simulation results show that DQN outperforms a dynamic programming algorithm adopting incorrect transition probabilities.

DDQN is used in [204] to solve a joint user association and channel allocation problem. Users adopt DDQN to find the optimal policy in terms of cell to connect to and channel to use, while taking into account a QoS constraint expressed in terms of minimum required SINR. The actions are tuples of selected cell and channel, while the state is a binary vector where the n-th element is 1 or 0 if the n-th user is or not satisfied (in terms of SINR) in the current configuration. Such a definition requires each user to know the satisfaction bit from the others: this is obtained via message passing from 1) users to the selected cell, and 2) across cells, to gather all the information from users. Finally, the information is transmitted back to the users. It is claimed that such a process results in negligible overhead, following the analysis in [205]. The user reward is defined as the difference between profits and costs. Profit is a function of the experienced data rate while cost depends on cell-specific prices and action-specific costs. Simulations are performed in a scenario with two macrocells, twenty-four small cells (eight pico and sixteen femtocells), a pool of thirty available channels of 180 kHz each, and fifty users requiring a minimum SINR of 5 dB. Results show that the convergence is obtained with a number of iterations decreasing as the DDQN learning rate increases from $\alpha = 0.001$ to $\alpha = 0.1$. The last value is thus used in the following evaluation, where DDQN is compared against DQN and Q-Learning, both with message passing, and shows higher learning speed and system capacity.

The same model is extended in [206], where dueling DDQN, referred to as D3QN, is proposed and adopted to solve the same selection-allocation problem. The analysis confirms the optimality of $\alpha = 0.1$, and also shows the impact of other DNN hyperparameters, e.g., the optimizer used to minimize the loss function and the DNN structure in terms of hidden layers and neurons. D3QN with message passing, referred to as D3QN(GS), is then compared with its counterpart without message passing, D3QN(SS), as well as with DQN, Q-Learning, a genetic algorithm, and a scheme based on maximum received signal power. It is shown that system capacity increases with the number of users; the absolute value depends on the adopted SINR constraint, in particular when the number of users increases from 40 to 50. The number of steps before convergence also increases with the number of users. Q-Learning based methods perform similarly, slightly in favour of D3QN(GS), while the genetic algorithm and the power-based scheme fail to achieve optimal policies when the number of users is high.

The work in [207] also adopts a dueling DQN architecture to solve RAT selection in a fully-distributed manner. The scenario includes a macrocell and several mmWave small cells. Users are part of a stochastic game and apply dueling DQN independently. Hence, they select an action (i.e., a cell), and get an individual reward (i.e., the experienced throughput). Then, they forward such rewards to the macrocell, that sends back the cumulative system throughput. For deriving their own selection policies, the users adopt the system throughput as reward signal, thus implicitly coordinating one another without exchanging information. In order to deal with partial state observability, and non-stationarity induced by the multi-agent setup, the hidden layers of DQN are composed of a RNN, as suggested by [208], which allows to better aggregate past information (e.g., previous observed states) in the decision making process. Finally, a so-called *hysteretic* mechanism is also adopted, as originally proposed in [209], where a higher learning rate is used to adjust the DQN weights when the TD error is positive, thus giving more importance to positive experiences. Simulation results in a scenario with one macrocell and three small cells, and six to thirteen users, demonstrate the effectiveness of the proposed approach, that outperforms the association scheme based on highest SINR, as well as a heuristic approach proposed in [210], in both static and mobile cases.

DQN is also proposed in [211] for symbiotic radio networks (SRNs) optimization. In this scenario, IoT devices parasitize cellular users for their own communications, aiming at spectrum-, energy-, and infrastructure-efficient communications. While a cell serves users in TDMA, a DQN derives the association between users and IoT devices; the latter aim to associate with the best possible users in order to reliably transmit their data using as power source the amount of power received and harvested from the cell. The optimal association would require real-time channel information while DQN exploits historical information. Two schemes are proposed, with the first making a joint decision for all IoT devices, and the second separately deciding for each device. Both schemes are deployed at the network side, with the second performed in separated computing units (one for each IoT device) in a same network entity. It is claimed that the second scheme can be independently applied by IoT devices with enough computing capabilities; this makes the analysis interesting in the context of the present work. Results show that the two schemes (which use $\epsilon = 0.2$ at the beginning, then ϵ decreases over time, $\alpha = 0.01$, and $\lambda = 0.3$) outperform random association and achieve performance near to the optimal policy.

C. RL-BASED SOLUTIONS WITH MAB MODELING

Stochastic games are multi-agent extensions of MDP and MAB models (Section V-C). A detailed analysis of MAB models and corresponding solving policies can be found in [39, Chapter 2], while an initial review of their application to networking problems is given in [212].

In its original form, the MAB model includes a learning agent having access to a set of *arms*, i.e., actions (or pure strategies), that provide an instantaneous reward (utility) upon selection. An introduction to stateful and stateless MAB models is provided below, along with a description of solving policies and their application to RAT selection, as also summarized in Table 5.

1) STATEFUL MAB

In stateful MAB, a state evolving upon selection can be associated with each arm and may be observed or not by the agent. Hence, the agent selects an arm at each time step, possibly observes the arm state transition (assumed Markovian), and gets an instantaneous reward, the distribution of which is stationary over time, depends on the state, and is unknown to the agent.[25] Once an arm is selected, the state of that arm, or the states of all arms, may change, leading to *restful* (or frozen) vs. *restless* Markovian MAB models, respectively.

The reward distribution in each state is unknown to the agent, which only observes a reward sample for the selected arm at each time step. Hence, without a learning process to guide its exploration vs. exploitation dilemma, the agent may select an arm leading to a lower reward with respect to the others, and in particular with respect to a possibly existing optimal arm. The latter is defined as the one having the highest average reward $\bar{u}^{\max} = \max_{a\in\mathcal{A}} \bar{u}(a)$, where a indicates a generic arm in the set $\mathcal{A}$, and $\bar{u}(a)$ represents the average reward of arm a. The average arm reward corresponds to the expected arm reward, since the reward distribution in each arm state is stationary. Hence, $\bar{u}(a) = \sum_{s_a\in\mathcal{S}_a}[\bar{u}(a|s_a) \times \Delta(u(a|s_a))]$, where s_a indicates a generic state for arm a within the set $\mathcal{S}_a$, $\bar{u}(a|s_a)$ is the average reward of arm a in state s_a, and $\Delta(u(a|s_a))$ is the distribution of the reward for arm a in state s_a. Notations similar to previous sections are adopted in order to highlight the aspects conceptually similar to GT. Hence, the MAB concept of reward is mapped onto the concept of utility and denoted as $u(\cdot)$, the arms are denoted as pure strategies, $a \in \mathcal{A}$, and the corresponding states are $s_a \in \mathcal{S}_a$ for all $a \in \mathcal{A}$.

On the one hand, keeping the optimal arm as a reference, MAB models and corresponding policies can be analyzed in terms of *regret*, that quantifies the cost of not selecting the optimal arm at each decision time. At time T, i.e., after T decisions, the agent cumulative regret is:

$$R_T^{\mathrm{MAB}} = T\bar{u}^{\max} - \mathbb{E}\Big[\sum_{t=1}^{T} \bar{u}(a_t|s_{a_t})\Big], \tag{29}$$

where $\bar{u}(a_t|s_{a_t})$ follows the definition given above and the subscript t specifies the arm selected at that time. The expectation reflects the state transition of each arm and the policy being used to select the arms. Such a policy would target the minimization of regret in (29).

On the other hand, similarly to MDPs, stateful MABs are also analyzed in terms of discounted utility, as discussed in Sections V-C and VI-B. This perspective is adopted in particular under the assumption of observable states and known state transitions for each arm. In this situation, so-called *indexing policies* have been proposed for solving Markovian MABs. These policies associate real scalar values, i.e., indexes, to each (arm, state) pair, indicating the reward that could be obtained by selecting a particular arm in that state. The arm with highest index is then selected at each decision step. The policy proposed by Gittins in [218] is optimal for frozen MABs, but the evaluation of Gittins indexes is not simple since it depends on multiple factors, including reward distributions. Hence, several algorithms to compute Gittins indexes have been proposed [219]. In case of restless bandits, Gittins and Whittle's [220] policies are asymptotically optimal under certain conditions, but their optimality does not hold in general. Further approximations for restless MABs with unknown dynamics have been found and discussed in [221] and references therein.

In the context of RAT selection, the work in [222] discusses the application of stateful MABs and Gittins policies for modeling and solving cellular handover, deferring the details for future work. Current handover schemes do not fully map into Gittins theory; hence, so-called *fractional* Gittins indices

[25] The description shows the similarity between MAB and MDP. However, in MAB, a different set of states and state transition distributions can be associated to each arm, while a MDP defines a unique state set for the *environment* surrounding the agent (at least in its most common form).

TABLE 5. Examples of application of MAB models and solutions to RAT selection.

Reference	MAB Model	Algorithm	Scenario Settings and Topology
[202]	Stochastic	ϵ-greedy, UCB1	$\|\mathcal{N}\|= 1$ $\|\mathcal{A}_{\text{tot}}\|= 8$ Overlapping: Generic cells
[213]	Stochastic	UCB1	$\|\mathcal{N}\|= [50:200]$ $\|\mathcal{A}_{\text{tot}}\|= [40:160]$ Nest: 1 Macrocell, varying Small cells
[110]	Stochastic	ϵ-greedy (benchmark), UCB1	$\|\mathcal{N}\|= [10^3, 10^5]$ $\|\mathcal{A}_{\text{tot}}\|= 5$ Overlapping: Small cells
[49]	Stochastic	UCB1 (benchmark)	$\|\mathcal{N}\|= 10$ $\|\mathcal{A}_{\text{tot}}\|= 3$ Nest: 1 Macrocell, 2 WLANs
[214]	Stochastic	ϵ-greedy (benchmark), UCB1 (benchmark and extended)	$\|\mathcal{N}\|= 1$ $\|\mathcal{A}_{\text{tot}}\|= 5$ Overlapping: Generic RATs
[101]	Adversarial	EXP3 and EXP4 (extended)	$\|\mathcal{N}\|= 6$ $\|\mathcal{A}_{\text{tot}}\|= [6, 12]$ Overlapping: Small cells
[111]	Adversarial (with sleeping arms)	ϵ-greedy (benchmark), EXP4	$\|\mathcal{N}\|= [50:70]$ $\|\mathcal{A}_{\text{tot}}\|= 5$ Overlapping: Small cells
[215]	Adversarial (periodic)	EXP3 (benchmark), EXP4 (extended)	$\|\mathcal{N}\|= 20$ $\|\mathcal{A}_{\text{tot}}\|= 3$ Overlapping: Generic RATs
[216]	Contextual	REXP3 [217] (benchmark), LinUCB	$\|\mathcal{N}\|= 1$ $\|\mathcal{A}_{\text{tot}}\|= 2$ Overlapping: Generic cells

are defined along with the procedures leading to optimal policies in case of restful MABs, which are also nearly optimal for restless MABs. It is demonstrated that handover based on thresholding schemes, e.g., CRE, leads to nearly optimal decisions after proper threshold setup.

2) STATELESS MAB

Stateless MABs simplify stateful MABs, since they do not have states; however, they may have non-stationary reward distributions. Hence, the average reward for an arm at time t is denoted as $\bar{u}_t(a)$ and equates to $\bar{u}(a)$ in case of stationarity. The regret can be written for instantaneous rewards, leading to the so-called *External* regret, as follows:

$$R_T^{\text{Ext}} = \max_{a\in\mathcal{A}} \mathbb{E}\Big[\sum_{t=1}^{T} u_t(a)\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} u_t(a_t)\Big], \tag{30}$$

Such a regret is commonly used for evaluating the performance of selection policies tailored for stateless MABs.

a: STOCHASTIC MODEL AND UPPER CONFIDENCE BOUND ALGORITHM (UCB1)

Stateless MABs with rewards following specific density functions (stationary or not) of given mean and variance are known as *Stochastic*, and are the most used MAB models, also in RAT selection problems. In the stationary case, the first term of (30) equates its counterpart in (29).

The pioneering work of Lai and Robbins [223] shows that the regret grows at least logarithmically over time. Then, optimal policies achieving this trend are derived for specific reward distributions, i.e., Bernoulli, Poisson, Gaussian, and Laplace.

The index policy referred to as UCB1, proposed by Auer *et al.* along with UCB2, UCB-Tuned, and UCB1-Normal extensions [224], shows logarithmic regret increase for any reward distribution defined on a bounded support.[26] According to (31), UCB1 evaluates an index for each available arm at each time step, denoted $i_T(a)$ for time T:

$$i_T(a) = \frac{\sum_{t=1}^{T_a} u_t(a)}{T_a} + \sqrt{\frac{2\ln T}{T_a}}. \tag{31}$$

The index is the sum of two terms: the current average reward and the size of the one-side confidence interval for it, that includes the true expected reward with overwhelming probability according to Chernoff-Hoeffding bounds [224]. In (31), T_a represents the number of times the agent has selected arm a, so that $T_a < T$. At time T, the arm with the highest index is then selected, as follows:

$$a_T = \arg\max_{a\in\mathcal{A}} i_T(a). \tag{32}$$

An application of UCB1 to cellular handover scenarios is proposed in [202], where it is first demonstrated that 3GPP protocol for handover between macrocells is a special case of the ϵ-greedy bandit algorithm, that is sub-optimal since it leads to a linear regret increase over time [39, Chapter 2].

[26] An extension of UCB1 to frozen MABs is proposed in [225].

Similar results are obtained for 3GPP handover schemes adopted in macro vs. small cells scenarios, such as CRE and the *sticky* protocol [69]. Finally, a novel, UCB1-based handover protocol is proposed, that takes into account handover costs in terms of overhead and transmission delay. The proposed protocol reduces the handover rate by introducing exploration in batches, that is, grouping time slots and forbidding handovers in there. Simulations are performed in a single-agent setup, considering a user having eight candidate cells, with throughput of each cell modeled according to a Gaussian distribution. Results show that UCB1 outperforms the 3GPP macrocell handover solution and leads to improvements over time due to learning. Moreover, UCB1 with batched exploration outperforms CRE and sticky protocols, reducing the number of handovers.

A similar handover scenario is analyzed in [213], that also introduces mmWave femtocells as candidate RATs. Due to peculiar propagation properties, the adoption of conventional handover mechanisms in mmWave HetNets may lead to ping-pong effects, high outage probability, and thus redundant handovers. The so-called SMART policy is introduced in order to determine the conditions triggering the handover, given both mmWave channel characteristics and QoS user requirements, and handle the selection of a cell among the candidates. SMART uses UCB1 in scenarios with low user density, while high density scenarios are modeled as a 0-1 integer programming optimization problem and solved by Lagrange dual decomposition with relaxation. SMART is compared against rate-based and SINR-based handover policies in a two-tier network comprising a 500 meters range macrocell and varying number of users and mmWave plus traditional femtocells. Simulation results show that SMART reduces the number of handovers up to 47% compared to the rate-based approach, leading to lower energy consumption and outage probability. In terms of experienced throughput, the rate-based approach outperforms the other schemes, but the improvement is limited compared to SMART and requires more frequent handovers. It is also shown that SMART is slightly more complex than the other algorithms, but presents an average signaling overhead similar to the rate-based scheme.

User association in cellular systems is also studied in [110], where an IoT scenario comprising a massive amount of devices transmitting in UL and embedding energy harvesting capabilities is considered. The multi-agent scenario is explicitly considered through the combination of so-called mean-field GT (MFGT) [226] and MAB, exploiting the *mean-field multi-armed bandit* game introduced by [227], [228]. Such a game relies on the *mean-field approximation*, that is applicable in case of large amounts of players. The approximation states that each player may consider the others as being stationary. Hence, the individual moves of players are irrelevant in terms of experienced utility, while the significant interaction is the one between each player and the mass of the others. In this context, the notions of equilibrium reported in Section IV-E are not practical, and the so-called mean-field equilibrium (MFE) is thus introduced. A MFE is a configuration where players split in portions adopting different strategies and such a split is stable over time. The work in [227], [228] demonstrates that MFEs exist under the assumption of Bernoulli-distributed utility for each arm. Moreover, adding further constraints, it is demonstrated that a unique MFE exists and is achieved by players adopting the so-called *mean-field dynamics*. These are defined so that, at each step, a player probabilistically incurs in either a regeneration process[27] or uses an arm selection policy that is common among users. The policy can be, for example, UCB1, that is tailored for single-agent scenarios but can be applied in multi-agent setups with similar implications already discussed for Q-Learning.

In [110], user association is thus modeled as a mean-field multi-armed bandit game and solved with UCB1, with no information exchange between users. Due to the need of a Bernoulli reward process for each strategy, the utility function is defined by taking the real value of the throughput experienced on the selected cell and turning it into a binary value, according to the probability of meeting a minimum required throughput. As stated by the authors, the Bernoulli reward process restricts the model applicability, and thus future work may focus on extending the analysis to other reward distributions. The convergence to the MFE is analyzed in a scenario with five small cells and 10^3 or 10^5 users. First, the mean-field dynamics perform better with 10^5 users, where the MFE is achieved with less fluctuations, reinforcing the idea that the mean-field approximation suits better games with a high number of players. Second, with 10^5 users and increased number of small cells (from three to seven), the convergence to the MFE is obtained in 50 to 100 iterations. Finally, with 10^3 users and three cells, the average number of successful transmissions and wasted energy efficiency (energy harvested but lost due to unsuccessful transmissions) are analyzed. UCB1 is compared against an optimal centralized solution (exhaustive search) and other distributed schemes: random association, ϵ-greedy, and *Explore-then-Commit*. In the latter, users select cells in a round-robin manner for a number of iterations (exploration); then, they connect to the cell that showed higher average number of successful transmissions during the exploration. Results show that UCB1 performs better than the other distributed schemes and consistently approaches the centralized solution.

UCB1 is used in [49] as a benchmark for three algorithms, referred to as online network selection (ONES), decoupled ONES (D-ONES), and virtual multiplexing ONES (VM-ONES). These three schemes are derived from modeling the selection problem as a continuous-time MAB (CT-MAB) [229]. In this case, playing an arm takes a random period of time, so the goal is to maximize the expected reward obtained in one unit time, i.e., to maximize the average

[27] In the RAT selection context, a regeneration step emulates the network dynamicity, i.e., a regenerating user leaves the selection game and a new user takes its place [110].

reward rate. Similarly, ONES, D-ONES, and VM-ONES aim at maximizing the reward rate defined as the ratio between users' QoE (in terms of MOS) and network access costs. QoE and costs are defined in terms of network state information (NSI), QoS, and traffic type (video, audio, or elastic). The algorithms are tested in a scenario including three available arms, i.e., two WLANs and one LTE macrocell, each one characterized by discrete models of delay, packet loss, and throughput. In a single-agent setup, the proposed schemes converge to the optimal selection policy. In terms of QoE reward rate, VM-ONES outperforms ONES, D-ONES, and two versions of UCB1 where QoE and costs are not considered but the selection is driven by either delay or throughput. In a multi-agent scenario, modeled by incorporating a congestion effect (network load increasing with the number of connected users), the algorithms converge to stable states, the game-theoretic nature of which is, however, not discussed. In scenarios with five to twenty users, VM-ONES achieves higher utility than ONES, D-ONES, Q-Learning [96], and BRD [43].

Measure-use MAB (muMAB) is proposed in [214] to better adapt MAB models to RAT selection. Classic MABs enable one possible action type in both exploration and exploitation phases, that is, to select an arm and collect the corresponding reward. However, a user may want to execute measurements on the candidate RATs (context retrieval), which may not require the connection to them. Moreover, such measurements (e.g., obtained via probing mechanisms and signaling) may be shorter in time than actual connections ($T_{\text{measure}} < T_{\text{use}}$), since T_{use} also include RAT switching procedures. Hence, muMAB differentiates between *measure* and *use* operations. In both cases, the user receives the reward of the RAT being selected as a feedback, but the reward translates into an actual performance gain only when the *use* operation is performed, since the RAT is exploited for data exchange only in that case. Both operations happen in batches, with the number of steps required to *measure* being lower than those spent to *use* a RAT. Two algorithms are designed and evaluated under the muMAB model: 1) measure-use-UCB1 (muUCB1) is derived from UCB1 and inherits the selection rule reported in (31)(32). muUCB1 performs a *use* operation when the selected arm corresponds to the one with the highest average reward estimate, otherwise it performs a *measure* operation; 2) measure with logarithmic interval (MLI) includes two phases. In phase 1, MLI performs consecutive *measure* operations on each arm, aiming at building reliable reward estimates. In phase 2, *use* operations are predominant apart for sporadic measurements vanishing over time. muUCB1 and MLI are compared against UCB1, ϵ-greedy and ϵ-decreasing algorithms, and the so-called price of knowledge and estimated reward (POKER) algorithm [230]. Several probability density functions are used for the reward of five RATs, including Bernoulli, positive-truncated Gaussian, and exponential distributions. Results show that performance depends on the adopted density function. muUCB1 and MLI guarantee best performance as the ratio between T_{use} and T_{measure} increases; muUCB1 is the best algorithm when the arms are characterized by similar average rewards, while MLI prevails when an arm is significantly more rewarding than all others.

b: ADVERSARIAL MODEL AND EXPONENTIAL-WEIGHT FOR EXPLORATION AND EXPLOITATION ALGORITHMS (EXP3/EXP4)

Another stateless MAB model known as *Adversarial* has been used in RAT selection [101], [111], [215]. In adversarial MAB, the reward does not follow a stochastic distribution but it is instead randomly decided by an *adversary* at each time step. The *Internal* regret is commonly used for performance analysis in an adversarial setup, defined as follows:

$$R_T^{\text{Int}} = \max_{a,a' \in \mathcal{A}} \sum_{t=1}^{T} p_t(a)[u_t(a) - u_t(a')], \tag{33}$$

where $p_t(a)$ is the probability of selecting arm a at time t and, differently from external regret, a pairwise comparison of arms is considered in R_T^{Int}. Note that, due to the adversarial setting, no assumption on the reward distribution is provided. The probabilities in (33) also suggest that adversarial MABs are usually solved in terms of mixed strategies, i.e., finding a probability distribution over the set of arms which minimizes R_T^{Int}.

Exponential-weight for exploration and exploitation (EXP3), and EXP3 using expert advice (EXP4), are common solving policies for adversarial MABs [231]. These algorithms evaluate mixed strategies and associate to each arm a selection probability proportional to the average experienced regret, weighted by an exponential function.

The work in [101] uses EXP3 and EXP4 to solve handover between small cells in an energy-efficient way. The proposed batched randomization with exponential weighting (BREW) algorithm performs batched exploration, reducing in this way unnecessary handovers. BREW relies on EXP3 and it is then enhanced via an EXP4-based learning strategy, referred to as ranking expert (RE). Both schemes counteract possibly missing feedback when selection is performed, and also deal with random activation and deactivation of small cells. The analysis is carried out in an environment including six or twelve small cells, symmetrically deployed at a distance of 80 meters around a small indoor area where several users randomly move. Results for a reference user show a significant energy consumption decrease with respect to 3GPP legacy handover solutions, and a better regret behaviour over time with respect to a genie-aided optimal scheme.

User association in a small cell network is analyzed in [111], where an adversarial MAB with *sleeping* arms is adopted. The sleeping feature models the possibility that the set of available arms is time-varying, i.e., not all and same arms are available during each selection step [232]. Small cells sleep when they harvest energy, and thus cannot serve

the users. Energy harvesting and consumption, channel quality, network traffic, and number of served users are unknown random variables that affect the probability of QoS satisfaction of users, i.e., successful DL transmissions. Users do not select cells simultaneously, so each cell has an ordered queue of users to be served, and deny service to those for which no energy is left. Given a selection step, served vs. denied users have $+1$ vs. 0 reward, respectively.

The adopted selection policy is the EXP4 version proposed in [231] for adversarial MABs with sleeping arms, that requires no information exchange across users. The only shared information is from cells, that announce their sleeping and activity periods via beacons. It is observed that a plain EXP4 algorithm could also be used, resulting in lower complexity but higher regret. In a network formed by five small cells and a number of users between 50 and 70, the analysis shows a similar behaviour for two reference users, both converging to the selection of cells maximizing their success probability. Similar results are obtained in a larger network composed by ten small cells, a number of users between 120 and 150, and observing 15 users. EXP4 is compared against a centralized optimal scheme, ϵ-greedy, ϵ-decreasing, Explore-then-Commit, random association, and two more centralized policies based on maximum received power and nearest cell. EXP4 shows improved performance at the cost of a reasonable complexity increase. Furthermore, results obtained in a small cell multihoming scenario show the flexibility of the proposed model.

An EXP4 extension is proposed in [215] for RAT selection. First, it is highlighted that bandit algorithms struggle in dynamic scenarios; as a matter of fact, they perform optimally in the stochastic case as they are usually designed to learn the average reward of each arm. However, learning the average reward may be sub-optimal if the scenario changes over time, even in a predictable manner, e.g., following periodic and repetitive patterns. In particular, RAT selection is modeled as a periodic scenario, since the available data rate on the RATs follows the behavior of users, which is repetitive and regulated by patterns over space and time. A periodic adversarial MAB is proposed and solved through periodic EXP4. The policy is evaluated by considering a periodic regret, which compares the obtained cumulative reward against the best possible periodic selection of arms. Moreover, periodic EXP4 exploits the repetitive structure of the targeted policy to reduce the computational complexity of EXP4. Simulations are performed in a scenario with 20 users and three RATs with periodic data rate and availability. Users perform one selection every minute over a period of two months, with network conditions having one-day periodicity. When connected to a RAT, users equally share the available rate. Periodic EXP4 is compared against EXP3 and a genie-aided scheme; it outperforms the former while approaching the latter under several settings, including discrete vs. continuous data rate change, and more realistic scenarios where imperfect knowledge of data rate pattern is modeled by adding Gaussian noise to exact data.

c: CONTEXTUAL MODEL AND LINEAR UPPER CONFIDENCE BOUND ALGORITHM (LinUCB)

In Contextual MAB, the agent observes at each time step the instantaneous reward and a multi-dimensional vector of *features* representing the selected arm (i.e., the context). The agent then derives a mapping between context and utility for each arm, and uses it to improve its future selections.

Algorithms developed for Markovian, Stochastic, and Adversarial MABs can be adapted and used in Contextual MABs, depending on the definition of reward. The linear UCB (LinUCB) algorithm proposed in [233] is widely used to solve Contextual MAB scenarios with stochastic utilities; given a feature vector observed at time t, LinUCB adopts ridge regression [234] to evaluate the expected utility for each available arm and, similarly to UCB1, selects the arm maximizing the estimated utility with a confidence bound.

LinUCB is used in [216] for solving RAT selection in intelligent transportation systems (ITSs). The scenario considers a train, embedded with sensors collecting and sending data to a gateway. The gateway then forwards the data to the cloud via LTE or Universal Mobile Telecommunications Service (UMTS). The ridge regression estimates the data rate over the two RATs based on several channel quality parameters.[28] Then, selection is performed that maximizes UL throughput. A slightly modified version of LinUCB is actually adopted, that adds a further confidence parameter to better guide selection under severe network congestion. This parameter is modeled as a rectified linear unit (ReLU) and enables the selection of a second RAT when the data rate obtained on the previous is much lower than the value estimated via regression. The parameter is periodically reset in order to mitigate selection bias due to perished information. The analysis is carried out using experimental data, including four separate measurement campaigns under mobility, and one campaign in a static setup. A commercial off-the-shelf (COTS) device is used as a gateway, embedded with both UMTS and LTE Subscriber Identity Module (SIM) cards, and generating User Datagram Protocol (UDP) traffic while collecting channel quality parameters. LinUCB is compared against a genie-aided algorithm, using a priori knowledge of the observed throughput for the candidate RATs, as well as four more approaches, that is, an EXP3 extension to non-stationary reward processes, presented in [217] and referred to as REXP3, and three algorithms using a different regression model each, namely linear, Bayesian, and support vector regression (SVR). These three use the regression during a training phase on a specific dataset and then select the RAT according to the values predicted by the trained model. Results show that LinUCB performance is affected by the data used in the training phase. By using a specific dataset for training, LinUCB performs better than the other algorithms and performs as well as the genie-aided scheme.

[28] Based on experiments, received signal strength indicator (RSSI) and energy per chip to power spectral density ratio ($\frac{E_c}{I_0}$) are used to estimate UMTS data rate, while RSSI and RSRP are adopted for LTE.

TABLE 6. Examples of application of RM, SLA and CODIPAS-RL to RAT selection.

Reference	Algorithm	Scenario Settings and Topology
[94]	RMv1	$\lvert\mathcal{N}\rvert = 100$ $\lvert\mathcal{A}_{\text{tot}}\rvert = 4$ Overlapping: Generic access nodes
[93]	RMv2, multiplicative weighted imitative CODIPAS-RL	$\lvert\mathcal{N}\rvert$ up to 20 $\lvert\mathcal{A}_{\text{tot}}\rvert$ up to 11 Nest: 1 Macrocell, up to 10 WLANs
[107]	SLA	$\lvert\mathcal{N}\rvert = 10$ $\lvert\mathcal{A}_{\text{tot}}\rvert = 2$ Overlapping: Generic access nodes
[235]	CODIPAS-RL (several algorithms)	$\lvert\mathcal{N}\rvert = [10, 10^4, 10^6]$ $\lvert\mathcal{A}_{\text{tot}}\rvert = 4$ Nest: 2 Macrocells, 2 WLANs
[236]	Bush-Morsteller and Boltzmann-Gibbs CODIPAS-RL	$\lvert\mathcal{N}\rvert = 2$ $\lvert\mathcal{A}_{\text{tot}}\rvert = 2$ Nest: 1 Macrocell (WiMAX), 1 WLAN

3) MULTI-AGENT MAB

From previous sections, several SARL-native algorithms developed under a MAB model, e.g., UCB1, EXP3, and EXP4, are adopted in multi-agent scenarios. On the one hand, this is nearly optimal from a signaling overhead perspective, since it removes the need for message exchange between agents. On the other hand, it may result in unsatisfactory results, since convergence to equilibria is not guaranteed and may require several explorations.

A step toward native multi-agent MAB is provided in [237], where a collaborative algorithm named Co-Bandit is proposed and applied to RAT selection modeled as a CG. In Co-Bandit, users probabilistically share information about their throughput with neighbors, and also forward delayed feedback received from others. Then, at each game step, users either explore or exploit; in the latter case, users follow the mixed strategy they are building based upon their own and neighbors' experience. Co-Bandit is tested against a full information algorithm and EXP3, in static and dynamic scenarios. In both cases, all users can hear each other; in the static case, the number of users (20) and available RATs (5) is constant, while it changes over time in the second case. The capacity of each RAT is selected considering WiFi and cellular performance, and this choice results in a unique NE for the game. Results show that Co-Bandit approaches the full information algorithm as users' cooperation increases, and rapidly converges to NE (a small gap is observed due to exploration). Co-Bandit outperforms EXP3 in terms of convergence speed and utility, and scales nicely as the number of users and RATs increase.

As observed in [212], [237], the application of MAB schemes to multi-agent scenarios leverages previous results proving possible convergence to NEs and CEs [51], [238], but still presents several open questions. Other recent results, that may also be adapted to RAT selection, can be found in [239]–[242], that extend MABs to multi-agent environments and propose algorithms leading to nearly logarithmically-increasing regrets. A systematic analysis of multi-agent MAB with respect to the issue of convergence from a game-theoretic perspective is, however, still missing.

D. OTHER APPROACHES

This section presents a further set of RL algorithms. The application to RAT selection games is discussed along with literature examples, also summarized in Table 6.

1) REGRET MATCHING (RM)

A valuable feature for a learning algorithm is to allow convergence to an extended set of equilibria, including CEs (Section VI-A). This can simplify the achievement of a stable and potentially efficient configuration. The two versions of RM [243] and [244] that are reviewed below converge to the set of CEs as $t \to +\infty$, that is a stable empirical distribution over pure strategy profiles is achieved.

An important aspect in RM is the notion of regret, that somehow extends the definition of regret in MAB models. As analyzed in [243], where the first RM algorithm, denoted RMv1 in the following, is introduced, each n-th player may be able to compute, at time t, the regret of not using the other available strategies in its set $\mathcal{A}_n$, given that it has selected strategy $a_{n,t}$. For all $a_n \neq a_{n,t} \in \mathcal{A}_n$, the set of regrets for player n is:

$$\varrho(a_n, a_{n,t}) = \max[0, D(a_n, a_{n,t})] \quad \text{for all } a_n \neq a_{n,t} \in \mathcal{A}_n, \tag{34}$$

where,

$$D(a_n, a_{n,t}) = \frac{1}{t} \sum_{\substack{\tau \leq t: \\ a_{n,\tau} = a_{n,t}}} [u_n(a_n, \boldsymbol{a}_{-n,\tau}) - u_n(a_{n,\tau}, \boldsymbol{a}_{-n,\tau}]. \tag{35}$$

At time t, the regret of not using a strategy $a_n \neq a_{n,t}$ is associated with the difference in the average utility that player n would experience if it had selected a_n every time in the past it has actually selected $a_{n,t}$. In RMv1, the set of regrets $\varrho(a_n, a_{n,t})$ are used to update the mixed strategy of player n at $t+1$, as reported in (36), as shown at the bottom of page 37.

In (36), μ is a large enough value to guarantee positive probabilities, and can be assumed fixed over time, e.g., $\mu > 2M_n(|\mathcal{A}_n| - 1)$, for all $n \in \mathcal{N}$, where M_n is the upper bound for the utility achievable by player n [243].

Several aspects can be discussed based on the RMv1 updating rule in (36) and regret evaluation in (34)(35). In particular, RMv1 can be directly applied to perfect and complete information games; in this case, player n observes the strategies selected by the others and is aware of how such strategies and its own affect utility. Then, it can evaluate utilities $u_n(a_n, \boldsymbol{a}_{-n,\tau})$, for all $a_n \neq a_{n,t} \in \mathcal{A}_n$ and $\tau \leq t : a_{n,\tau} = a_{n,t}$. In order to evaluate $u_n(a_{n,\tau}, \boldsymbol{a}_{-n,\tau})$, a player does not need to observe the others, since this is the utility experienced upon selecting $a_{n,\tau}$ and does not require perfect and/or complete information for its evaluation.

Slight modifications of (34)-(36) are proposed in [244] and lead to a second RM algorithm (RMv2), that can be used in imperfect and incomplete information games while preserving the convergence to the set of CEs. Instead of relying on the actual regrets, player n estimates them by evaluating the set of $\varrho^{\text{est}}(a_n, a_{n,t}) = \max[0, D^{\text{est}}(a_n, a_{n,t})]$, for all $a_n \neq a_{n,t} \in \mathcal{A}_n$, with D^{est} as in (37), as shown at the bottom of the next page.

In this case, player n relies on the past experienced utilities of selecting either a_n or $a_{n,t}$ for the evaluation of both terms in (37), ultimately requiring neither the knowledge of the strategies played by the others nor the impact on the utility. $p_{n,\tau}(a_n)$ indicates the probability of player n of selecting strategy a_n at game step τ. The updating rule for RMv2 is then given in (38), as shown at the bottom of the next page [244], where $0 < \delta < 1$, $0 \leq \gamma < \frac{1}{4}$, and μ follows (36). Further details on the choice of these parameters are provided in [244].

In the context of RAT selection, RMv1 is applied in [94], where a simplified PF model with RAT-specific fixed connection price is considered as utility. In order to cope with the lack of complete information, each RAT broadcasts the missing information, enabling users to estimate the utility achievable over the non-selected RATs, and in turn the regret of not selecting them. The scenario includes 100 users and 4 candidate RATs; RMv1 shows convergence to a CE near to the SO profile in about 15 iterations. RMv2 is considered in [93] to solve an imperfect and incomplete information game with noisy TF and PF utilities. A preliminary evaluation on a 2-user 2-RAT game shows that RMv2 converges after many iterations (about 6×10^3) and switchings across the RATs (about 400) to a CE that is neither user- nor social-optimal. Hence, network assistance is proposed to cope with low performance. The access nodes broadcast denoised information that are used by users in a version of RMv2 with a slightly different updating rule. Such a scheme shows better performance with respect to the original algorithm and the BRD solution of [44] (Section VI-A1).

2) STOCHASTIC LEARNING AUTOMATA (SLA)

Proposed in [245], SLA is a RL algorithm converging to PNEs in PGs (Section V-A). SLA agents exploit their own instantaneous utility as reinforcement signal and adjust the adopted mixed strategy over time, as given in (39), as shown at the bottom of the next page, where $\tilde{u}_{n,t}$ is the normalized utility and $0 < \alpha < 1$ is the learning rate.[29] The updating rule in (39) derives from (14), and states that a high experienced utility results in a high selection probability in the next step. SLA converges to NEs if α is sufficiently small [245].

SLA is used in [107] to solve RAT selection in a cognitive radio (CR) scenario. The so-called secondary users (SUs) can connect to several primary networks for their data traffic, but the availability of transmission channels depends on the time-varying demands of primary users (PUs). The utility function follows the PF definition in (2b), and depends on the amount and capacity of available channels. The selection game is an ordinal PG (OPG), that is, a relaxed version of EPG still presenting PNEs [135]. SLA is evaluated in a scenario with ten SUs and two RATs with three channels each, where the channels have predefined capacities. It is shown that users adopting a random selection in the beginning and then SLA converge to pure strategies in about 100 vs. 300 steps with $\alpha = 0.5$ and $\alpha = 0.2$, respectively. However, the strategy profile achieved with $\alpha = 0.5$ is not a PNE, since a unilateral deviation of a SU leads the SU to a higher throughput. SLA is compared against two other approaches: 1) a scheme where users select the RAT with the best per-channel throughput and 2) a centralized scheme based on exhaustive search. With randomly distributed SUs, SLA outperforms the first scheme, but is far from the second in terms of system throughput (defined as the sum of per-user throughputs). However, compared to the centralized scheme, SLA converges to a solution with higher fairness across users, evaluated via Jain's index (see Section VII). In a Nest network where the SUs select between an indoor small cell and a macrocell located far apart, SLA converges to a PNE that is optimal in terms of system throughput and fairness.

3) COMBINED FULLY DISTRIBUTED PAYOFF AND STRATEGY REINFORCEMENT LEARNING (CODIPAS-RL)

Similarly to SLA, the approaches under the CODIPAS-RL framework follow the basic updating rule in (14), and differ depending on how the agent utility is taken into account when the selection probabilities are updated at $t \rightarrow t + 1$ [59]. Among others, CODIPAS-RL include so-called *Bush-Morsteller* [246], *Boltzmann-Gibbs*, and *multiplicative weighted imitative* algorithms. The updating rule for these algorithms can be found in [235], where the use of the entire framework is proposed for RAT selection. Beyond the three above methods, the paper introduces other schemes with similar updating rules and discusses the assumptions on the game structure under which the schemes converge to equilibria. The overall model allows users to adopt different CODIPAS-RL schemes with variable learning rates. Users can also

[29] A constant and player-agnostic α is assumed in (39). This is in agreement with [245], where SLA was originally proposed, and [107], where SLA is used for RAT selection. However, α may be different across players and time-decreasing, as discussed for Q-Learning in Section VI-B1.

a) switch across schemes during the game, thus performing *heterogeneous learning*, and b) switch between active and sleep modes (i.e., participate in the selection game only when active). The use in a 2-operator 4G / WiFi network is proposed and tested by network simulation. Results show convergence to stable configurations in different scenarios, including underloaded vs. congested RATs and different service costs.

Bush-Morsteller and Boltzmann-Gibbs CODIPAS-RL are compared in a WiMAX/WLAN scenario in [236], where the latter shows faster convergence. Moreover, Multiplicative Weighted Imitative CODIPAS-RL is used to benchmark the network-assisted solution in [93] (Section VI-D1), showing low performance in terms of convergence speed, per-user switchings, and system utility compared to other learning schemes, that however exploit more information and network assistance. It is interesting to observe that the algorithm achieves an equilibrium showing high fairness across users; this is explained by considering that its learning dynamics are similar to the ones of RD, as discussed in [247].

VII. PERFORMANCE INDICATORS

The analysis of user-centric RAT selection in a joint GT-MAL approach is a challenging task. The case can be modeled and solved in different ways and ultimately analyzed based on multiple performance indicators. The heterogeneity of such indicators highlights that a multi-faceted evaluation framework is needed in order to exhaustively and reliably compare different proposed solutions. This section discusses commonly used indicators, by grouping them in three main categories: *Convergence*, *Efficiency*, and *Fairness*.

A. CONVERGENCE

This category includes indicators of when and how a proposed learning scheme converges to a game-theoretic equilibrium. The following metrics are commonly used:

- Evolution over time of the number (or percentage) of users connected to each RAT;
- Evolution over time of the probability for a user to select each RAT;
- Number of iterations, i.e., game steps, leading to an equilibrium;
- Number of switchings across RATs leading to an equilibrium (for each user or by averaging across users).

The stability over time of the first two metrics suggests a possible convergence to an equilibrium. For example, when the second metric is independently evaluated for each user, it may show convergence to a NE. The last two metrics indicate instead how the proposed scheme moves toward the equilibrium. As observed in previous sections, the dynamicity of RAT selection requires a rather fast convergence to a stable strategy profile, due to possible sudden perturbations provoked by the arrival of new users for example, or the detection of new candidate RATs, mobility, etc. Such changes may invalidate the equilibrium, and require a new learning process. Moreover, repeated switching across RATs results in additional costs in terms of device energy consumption and connection prices; hence, an excessive number of switchings should be discouraged.

B. EFFICIENCY

This category includes the metrics on utility performance obtained during the learning process and/or at equilibrium. RAT selection games may present several equilibria, that may be different in terms of per user and overall utility. Moreover, the equilibria may perform poorly when compared to PO and SO strategy profiles.

Initial investigations adopting complete information PGs and CGs [94], [102], [112] use two specific performance indicators to assess the efficiency of game equilibria, i.e., price of anarchy (PoA) and price of stability (PoS).

Definition 7 Price of Anarchy (Stability) – PoA (PoS) [248], [249]: Given a finite strategic game and a welfare

$$\text{RMv1:}\quad p_{n,t+1}(a_n) = \begin{cases} \dfrac{1}{\mu}\varrho(a_n, a_{n,t}) & \text{for all } a_n \neq a_{n,t} \in \mathcal{A}_n \\ 1 - \displaystyle\sum_{\substack{a_n \in \mathcal{A}_n: \\ a_n \neq a_{n,t}}} p_{n,t+1}(a_n) & a_n = a_{n,t} \end{cases} \tag{36}$$

$$D^{\text{est}}(a_n, a_{n,t}) = \frac{1}{t} \sum_{\substack{\tau \leq t: \\ a_{n,\tau} = a_n}} \frac{p_{n,\tau}(a_n)}{p_{n,\tau}(a_{n,t})} u_n(a_n, \boldsymbol{a}_{-n,\tau}) - \frac{1}{t} \sum_{\substack{\tau \leq t: \\ a_{n,\tau} = a_{n,t}}} u_n(a_{n,t}, \boldsymbol{a}_{-n,\tau}). \tag{37}$$

$$\text{RMv2:}\quad p_{n,t+1}(a_n) = \begin{cases} \left(1 - \dfrac{\delta}{t^\gamma}\right) \min\left\{ \max\left\{0, \dfrac{D^{\text{est}}(a_n, a_{n,t})}{\mu}\right\}, \dfrac{1}{|\mathcal{A}_n| - 1}\right\} + \dfrac{\delta}{t^\gamma |\mathcal{A}_n|} & \text{for all } a_n \neq a_{n,t} \in \mathcal{A}_n \\ 1 - \displaystyle\sum_{\substack{a_n \in \mathcal{A}_n: \\ a_n \neq a_{n,t}}} p_{n,t+1}(a_n) & a_n = a_{n,t} \end{cases} \tag{38}$$

$$\text{SLA:}\quad p_{n,t+1}(a_n) = \begin{cases} p_{n,t}(a_n) + \alpha \tilde{u}_{n,t}[1 - p_{n,t}(a_n)] & a_n = a_{n,t} \\ p_{n,t}(a_n) - \alpha \tilde{u}_{n,t} p_{n,t}(a_n) & \text{for all } a_n \neq a_{n,t} \in \mathcal{A}_n \end{cases} \tag{39}$$

function $\mathsf{w}: \mathcal{A}_{\text{tot}} \to \mathbb{R}$, the PoA (PoS) is defined as the ratio between the optimal solution and the worst (best) NE in terms of welfare function w. Considering pure strategy profiles and PNEs,[30] PoA and PoS are defined as follows:

$$\mathsf{PoA} := \frac{\max\limits_{\boldsymbol{a}\in\mathcal{A}_{\text{tot}}} \mathsf{w}(\boldsymbol{a})}{\min\limits_{\boldsymbol{a}^{\text{NE}}\in\mathcal{A}_{\text{tot}}^{\text{NE}}} \mathsf{w}(\boldsymbol{a}^{\text{NE}})}, \tag{40}$$

$$\mathsf{PoS} := \frac{\max\limits_{\boldsymbol{a}\in\mathcal{A}_{\text{tot}}} \mathsf{w}(\boldsymbol{a})}{\max\limits_{\boldsymbol{a}^{\text{NE}}\in\mathcal{A}_{\text{tot}}^{\text{NE}}} \mathsf{w}(\boldsymbol{a}^{\text{NE}})}, \tag{41}$$

where $\mathcal{A}_{\text{tot}}^{\text{NE}}$ represents the set of game PNEs and $1 \leq \mathsf{PoS} \leq \mathsf{PoA}$. Several welfare functions can be used; among others, the sum of utilities across players is commonly adopted, i.e., $\mathsf{w}(\boldsymbol{a}) = \sum_{n\in\mathcal{N}} u_n(\boldsymbol{a})$. In this case, PoA and PoS reflect a comparison between the SO profile and worst and best PNE. The calculation of PoA and PoS requires the evaluation of the set of equilibria, that can be done, for example, by solving the system of non-deviation inequalities [112]. With a similar scope, [44] reports the average number of equilibria for the proposed game and the fraction of them being PO.

Recent work is more focused on learning under imperfect and incomplete information assumptions and thus provides further insights on learning efficiency. The following metrics are commonly adopted:

- Per user utility over time;
- System utility over time, evaluated by averaging (or summing up) the user utility at each iteration.

A few papers adopting MAB models use regret to analyze learning efficiency [101], [214]. Algorithms are also compared in terms of signaling overhead in [93].

C. FAIRNESS

This category includes metrics that quantify to what extent the proposed scheme promotes utility fairness between users. Besides evolutionary games solved via RD, the equilibria do not have fairness constraints per se, and this may be a drawback for specific selection scenarios, e.g., offloading. When fairness is explicitly considered, as for example in [46], [93], [102], among others, the Jain's index is used [250]:

$$J := \frac{(\sum_{n=1}^{N} \bar{u}_n)^2}{N \times (\sum_{n=1}^{N} \bar{u}_n^2)}, \tag{42}$$

where $\bar{u}_n$ denotes the average utility experienced by user n, that is, the average throughput over time.

VIII. OPEN CHALLENGES

A. GT AND MAL EVOLUTION

After many years of development and application to several societal contexts, GT is undoubtedly nowadays a mature framework, while MAL is more recent and still under investigation, particularly in its DRL form. Hence, while many connections between GT and MAL have been already discovered, many others still need to be unveiled. Moreover, novel MAL algorithms will most likely be proposed in the next years, and the analysis of the relationship with GT is key for their validation [36], [37].

User-centric RAT selection is a relevant use case for both theories. The analysis of current literature shows that SARL-native algorithms have been commonly adopted for solving RAT selection games. Further work is thus needed toward the analysis of schemes tailored for multi-agent scenarios. New GT models and MAL algorithms will provide improvement of RAT selection schemes, leading to benefits toward smart connectivity in next-generation systems and networks.

A recent example is federated learning (FL), that is a MAL framework that takes advantage of an increased computational capability of end devices. In FL, devices collect and use their own data to train a local ML model. Model parameters are then transmitted to a higher-layer server, either located in the cloud or at the network edge [251]. Hence, devices do not send large amounts of data to the servers, but instead only transmit processed parameters, ultimately enhancing data privacy and avoiding communication overhead. FL application to communication networks is surveyed in [252], where FL mapping onto edge mobile networks is also highlighted. Besides providing details on FL and open challenges, FL applications to edge networking scenarios are also reported, including cell association.

A FL-based approach is used in [253] for enabling the sharing of local association policies among neighbor users. The scenario is modeled as a mean-field game where it is also assumed that users adopt an imitation mechanism. Users exploit local conditions and policies of its neighbors to derive their own DQL-based policy. The assumption is that neighbors face similar conditions during their selections. The work shows the advantage of collaborative learning.

In parallel, DRL-related research is increasingly focused on multi-agent scenarios, thus enabling possible applications to RAT selection. As selected examples, an actor-critic DQN is proposed in [254] to converge to NEs in stochastic games; the mean-field case is investigated in [255]. Focusing on deep policy gradient methods rather than DQL, [256] proposes a multi-agent extension of deterministic schemes [257].

B. COMMUNICATION SYSTEMS EVOLUTION

Along with GT and MAL, communication systems are also evolving. Their enhancement leads to new challenges to address, as clear nowadays with the advent of 5G. Network evolution will continue beyond 5G with increased use of artificial intelligence (AI) [258], [259].

From a RAT selection perspective, it is important to keep aligning modeling assumptions. In particular, *heterogeneity and massive connectivity*, in terms of new RATs and large amount of users, require the extension of modeling. MFGT models seem a viable solution to handle ultra-dense

[30] Definition 7 can be straightforwardly extended to MNEs and CEs.

scenarios, but the use in combination with MAL algorithms is still in its infancy [110], [253], [255].

Moreover, *novel scenarios and services*, such as eMBB, mMTC, and URLLC for 5G, require to better cope with user heterogeneity. Coexisting users may have different goals and thus their selection strategies could be driven by a different utility. The latter may fall apart from traditionally used throughput-based definitions, that seem directly applicable to eMBB scenarios only. Hence, new definitions are needed based on QoS indicators of service reliability and energy efficiency/consumption (mMTC) and reliability/latency (URLLC) [260]. A similar observation can be done for novel vehicular communications, for which initial GT-based RAT selection approaches can be found in [261], [262]. Even when they have similar goals, users may still show different capabilities in terms of context observation and computation, eventually adopting different learning schemes. Besides initial work under CODIPAS-RL [235], heterogeneous learning is still marginally investigated in RAT selection and thus represents an extension for future work.

As mentioned while introducing FL, mobile edge networks and mobile (more recently, multi-access) edge computing (MEC) are becoming of extreme interest for the design and development of next generation systems. MEC enables the placement of computational servers at the network edges, e.g., co-located with radio access nodes. Such servers host different network functions nearer to users, thus reducing latency, congestion, and communication overhead with respect to cloud architectures. MEC ultimately enables the provisioning of services requiring extreme low latency and high computational capabilities (e.g., tactile Internet, self-driving cars, and IoT analytics, to mention a few). A survey on MEC concepts and applications to 5G and beyond networks can be found in [263], while a survey connecting GT and MEC is given in [264]. In the context of RAT selection, a MEC instance may allow users to offload some of their energy- and cost-demanding learning tasks to the edge servers (e.g., train and maintain a D*Q*N), and simplify network-assisted selection schemes. MEC-enabled access nodes may have sufficient resources to evaluate specific user conditions and help a fast discovery of optimal selection policies. Moreover, information between users and servers is exchanged on shorter paths, allowing fast reaction in dynamic scenarios (e.g., highly mobile users).

Finally, *dynamic resource sharing and coexistence* among different RATs is a rapidly evolving paradigm, but still marginally investigated in selection scenarios. An example is in the spectrum domain, where spectrum sharing mechanisms have been proposed in cellular standards, e.g., enabling LTE to operate in unlicensed bands and competing with WiFi for spectrum resources [15], [265], [266]. Such a paradigm is being recently extended to 5G NR, which may compete in the 60 GHz spectrum against IEEE 802.11ad/ay, i.e., Wireless Gigabit (WiGig) [267], [268]. From a modeling perspective, a utility only affected by users connecting to the same RAT (e.g., as for CGs) is not representative of spectrum sharing scenarios, since the selection of a different RAT may still affect the utility, e.g., in terms of interference.

C. COMPARISON ACROSS SIMILAR MECHANISMS

As reviewed in Section III-A, multiple approaches for enabling RAT selection exist in practice. While several approaches are proposed to function at the radio layer, other approaches work at the transport layer, e.g., via MPTCP and MPQUIC protocols.

The core component of a MP transport protocol is the *scheduler*, that decides how to redirect data on the available RATs (*paths*) at either flow or packet level. Several policies driving scheduling decisions have been proposed, ranging from simple path-agnostic round robin (RR) to more sophisticated mechanisms that consider path status and characteristics. Among them, the so-called minimum round trip time (minRTT) scheme redirects data on the path with lowest delay and is the default scheduler for MPTCP and MPQUIC [269]. Other schedulers have been proposed to deal with out-of-order delivery and blocking, as reported in [270]. An experimental comparison of MPTCP schedulers can be found in [271].

On the one hand, recent work has initiated the modeling of MP scheduling as a decision-making problem, ultimately leveraging RL-based solutions. In particular, reinforcement learning based scheduler (ReLeS) in [272] adopts D*Q*N in order to find optimal scheduling policies; Peekaboo [270] models the scheduling task as a contextual MAB and employs LinUCB and a stochastic adjustment in order to derive a probabilistic policy. Modified-Peekaboo is proposed in [273], in order to better deal with the high dynamicity of 5G mmWave access. On the other hand, recent standardization, such as the ATSSS architecture being proposed by 3GPP for 5G and beyond, is moving toward cross-layer solutions, aiming at full network interoperability. Further analyses are thus needed to highlight the best approaches to use in different network scenarios (e.g., *when and why is it better to adopt network interoperability schemes at radio and/or transport layers?*), ultimately leading to improved and convergent solutions.

D. DATA-DRIVEN MODELING, ANALYSIS, AND TESTING

The overview of RAT selection literature shows an ongoing transition from pure theoretical modeling (e.g., under perfect and complete information assumptions) to more practical implementation of learning schemes (e.g., assuming imperfect and incomplete information). More recent investigations have also proposed empirical analyses due to higher availability of data collected in experimental open platforms, testbeds, and large-scale measurement campaigns [93], [181], [261]. On this aspect, differently from RAT selection, MP scheduling solutions are most often analyzed via experiments in real scenarios, also thanks to the possibility of using existing software implementations for several MP protocols. For example, MPTCP implementation in the Linux kernel is available

online.[31] Such implementations can be embedded with new components, e.g., a new scheduler, ultimately enabling the validation in real networks. It is thus clear the need for data-driven RAT selection analyses, aiming at deriving more realistic models in terms of utility, and test proposed methods in real systems and scenarios.

IX. CONCLUSION

This paper provides a unified reference in the context of user-centric RAT selection regarding most commonly adopted GT models and MAL algorithms, and highlight how these map onto RAT selection scenarios. The GT and MAL overview is complemented by a discussion on the assumptions commonly made in RAT selection in terms of utility function, network topology, and key performance indicators.

The review provides a comparative literature analysis, and emphasizes modeling trends and achievements. Open challenges and future work are also discussed, based on surveyed literature, recent advances in GT, MAL, as well as ongoing standardization activities related to RAT selection.

The present work ultimately provides a reference for ongoing and future research activities, toward designing high-performing user-centric RAT selection schemes.

APPENDIX
ACRONYMS

3GPP	3rd generation partnership project
4/5G	4th/5th generation
A3C	Asynchronous advantage actor-critic
ABC	Always best connected
ACR	Available capacity ratio
AE	Auto-encoder
AHP	Analytical hierarchy process
AI	Artificial intelligence
ANDSF	Access Network Discovery and Selection Function
AP	Access point
ATSSS	Access Traffic Steering, Switching, and Splitting
BER	Bit error rate
(a/s)BR(D)	(Asynchronous/simultaneous) best response (dynamics)
BS	Base station
BP	Boltzmann procedure
BREW	Batched randomization with exponential weighting
CE	Correlated equilibrium
(W)CDMA	(Wideband) code division multiple access
(u/w)CG	(Unweighted/weighted) congestion game
CMT	Concurrent Multipath Transfer
CNP	Coupled network pair
CODIPAS-RL	Combined fully distributed payoff and strategy reinforcement learning
CoMP	Coordinated multi-point
COTS	Commercial off-the-shelf
CRE	Cell range expansion
D3QN	Dueling double deep Q-Network
DC	Dual connectivity
DDPG	Deep deterministic policy gradient
DQL	Deep Q-Learning
(D)DQN	(Double) deep Q-Network
DL/UL	Downlink/uplink
eMBB	Enhanced mobile broadband
eNB	evolved Node B
ESS	Evolutionary stable strategy
EXP3	Exponential-weight algorithm for exploration and exploitation
EXP4	Exponential-weight algorithm for exploration and exploitation using expert advice
FIP	Finite improvement path
FL	Federated learning
FP	Fictitious play
GAN	Generative adversarial network
GRA	Gray relational analysis
(e/MF)GT	(Evolutionary/mean-field) game theory
HetNets	Heterogeneous networks
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IoT	Internet of Things
IP	Internet Protocol
ITS	Intelligent transportation system
***k*NN**	*k*-nearest neighbors
(E-)LIA	(Enhanced-) local improvement algorithm
LTE(-A)	Long-Term Evolution (-Advanced)
LWA(AP)	LTE-WLAN Aggregation (Adaptation Protocol)
LWIP(EP)	LTE-WLAN radio level integration with Internet Protocol security tunnel (Extension Protocol)
(CT/mu)MAB	(Continuous time/measure-use) multi-armed bandit
MAC	Medium Access Control
MADM	Multiple attribute decision making
MA(R)L	Multi-agent (reinforcement) learning
MCDM	Multiple criteria decision making
MCTS	Monte Carlo tree search
(PO/S)MDP	(Partially observable/semi-) Markov decision process
MEC	Mobile (or multi-access) edge computing
MEW	Multiplicative exponential weighting
MFE	Mean field equilibrium
MIH	Media Independent Handover
ML	Machine learning
MLI	Measure with logarithmic interval
(m)MTC	(Massive) machine type communications
mmWave	Millimeter-wave
MOS	Mean opinion score

[31] "MultiPath TCP - Linux kernel implementation", http://www.multipath-tcp.org, Accessed on: May 2021.

MP	Multipath
N3IWF	Non-3GPP Inter-Working Function
(P/M)NE	(Pure/mixed) Nash equilibrium
(C/D/F/R)NN	(Convolutional/deep/feedforward/recurrent) neural network
NR	New Radio
NSA	Non-Standalone
NSI	Network state information
OFDMA	Orthogonal frequency division multiple access
(D/VM)ONES	(Decoupled/virtual multiplexing) online network selection
PDCP	Packet Data Convergence Protocol
(E/O)PG	(Exact/ordinal) potential game
PO	Pareto-optimal
PoA/PoS	Price of anarchy/stability
POKER	Price of knowledge and estimated reward
PPP	Poisson point process
PSR	Packet success ratio
QBNS	Q-Learning based network selection
QoS/QoE	Quality of service/experience
RAN	Radio access network
RAT	Radio access technology
RD	Replicator dynamics
RE	Ranking expert
ReLeS	Reinforcement learning based scheduler
ReLU	Rectified linear unit
RFEQG	Random forest enhanced Q-Learning with game
(D)RL	(Deep) reinforcement learning
RM	Regret matching
RR	Round robin
RSR(P/Q)	Reference signal received (power/quality)
RSS(I)	Received signal strength (indicator)
(min)RTT	minimum round trip time
SA(R)L	Single-agent (reinforcement) learning
SAW	Simple additive weighting
SCTP	Stream Control Transmission Protocol
SIM	Subscriber Identity Module
SINR	Signal to interference plus noise ratio
SLA	Stochastic learning automata
SO	Social-optimum
SPE	Subgame perfect Nash equilibrium
SRN	Symbiotic radio network
SVR	Support vector regression
TCP	Transport Control Protocol
TD	Temporal-difference
TDMA	Time division multiple access
TF/PF	Throughput fair / proportional fair
TOPSIS	Technique for order preference by similarity to ideal solution
UAV	Unmanned aerial vehicle
(Lin)UCB	(Linear) upper confidence bound
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications Service
URLLC	Ultra reliable low latency communications
VDBE	Value-difference based exploration
VLC	Visible light communication
WBAN	Wireless boby area network
WiGig	Wireless Gigabit
WLAN	Wireless local area network
WMAN	Wireless metropolitan area network
WiMAX	Worldwide Interoperability for Microwave Access
WPAN	Wireless personal area network
WWAN	Wireless wide area network

REFERENCES

[1] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 90–96, Feb. 2014.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[3] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Commun.*, vol. 10, no. 1, pp. 49–55, Feb. 2003.

[4] S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S.-P. Yeh, and S. Talwar, "Intelligent access network selection in converged multi-radio heterogeneous networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 86–96, Dec. 2014.

[5] R. Trestian, O. Ormond, and G.-M. Muntean, "Game theory-based network selection: Solutions and challenges," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 1212–1231, 4th Quart., 2012.

[6] L. Wang and G.-S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks—A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 271–292, 1st Quart., 2013.

[7] A. Ahmed, L. M. Boulahia, and D. Gaiti, "Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 776–811, 2nd Quart., 2014.

[8] F. Rebecchi, M. Dias de Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti, "Data offloading techniques in cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 580–603, 2nd Quart., 2015.

[9] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.

[10] M. Wang, J. Chen, E. Aryafar, and M. Chiang, "A survey of client-controlled HetNets for 5G," *IEEE Access*, vol. 5, pp. 2842–2854, Nov. 2017.

[11] G. Dandachi, S. E. Elayoubi, T. Chahed, and N. Chendeb, "Network-centric versus user-centric multihoming strategies in LTE/WiFi networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4188–4199, May 2017.

[12] P. Marsch, B. Raaf, A. Szufarska, P. Mogensen, H. Guan, M. Farber, S. Redana, K. Pedersen, and T. Kolding, "Future mobile communication networks: Challenges in the design and operation," *IEEE Veh. Technol. Mag.*, vol. 7, no. 1, pp. 16–23, Mar. 2012.

[13] H. Elsawy, E. Hossain, and D. Kim, "HetNets with cognitive small cells: User offloading and distributed channel access techniques," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 28–36, Jun. 2013.

[14] M. Alam, D. Yang, K. Huq, F. Saghezchi, S. Mumtaz, and J. Rodriguez, "Towards 5G: Context aware resource allocation for energy saving," *J. Signal Process. Syst.*, vol. 83, no. 2, pp. 279–291, May 2016.

[15] G. Caso, L. De Nardis, and M.-G. Di Benedetto, "Toward context-aware dynamic spectrum management for 5G," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 38–43, Oct. 2017.

[16] D. D. Nguyen, H. X. Nguyen, and L. B. White, "Evaluating performance of RAT selection algorithms for 5G hetnets," *IEEE Access*, vol. 6, pp. 61212–61222, Oct. 2018.

[17] M. Chiang, S. Ha, F. Risso, T. Zhang, and I. Chih-Lin, "Clarifying fog computing and networking: 10 questions and answers," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 18–20, Apr. 2017.
[18] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.
[19] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, Mar. 2013.
[20] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100–1113, Jun. 2014.
[21] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1040–1054, Jun. 2015.
[22] Q. Wu, M. Tao, and W. Chen, "Joint Tx/Rx energy-efficient scheduling in multi-radio wireless networks: A divide-and-conquer approach," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2727–2740, Apr. 2016.
[23] T. Z. Oo, N. H. Tran, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Offloading in HetNet: A coordination of interference mitigation, user association, and resource allocation," *IEEE Trans. Mobile Comput.*, vol. 16, no. 8, pp. 2276–2291, Aug. 2017.
[24] Z. Chkirbene, A. Awad, A. Mohamed, A. Erbad, and M. Guizani, "Deep reinforcement learning for network selection over heterogeneous health systems," *IEEE Trans. Netw. Sci. Eng.*, early access, p. 1, Feb. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9351659
[25] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.
[26] M. J. Osborne and A. Rubinstein, *A Course Game Theory*. Cambridge, MA, USA: MIT Press, 1994.
[27] J. M. Vidal, *Fundamentals of Multiagent Systems With NetLogo Examples*. Citeseer, 2006. [Online]. Available: http://jmvidal.cse.sc.edu/papers/mas.pdf
[28] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, vol. 2. Cambridge, MA, USA: MIT Press, 1998.
[29] H. P. Young, *Strategic Learning and its Limits*. London, U.K.: Oxford Univ. Press, 2004.
[30] L. Blumrosen and N. Nisan, *Algorithmic Game Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[31] S. Lasaulce and H. Tembine, *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*. New York, NY, USA: Academic, 2011.
[32] G. Bacci, S. Lasaulce, W. Saad, and L. Sanguinetti, "Game theory for networks: A tutorial on game-theoretic tools for emerging signal processing applications," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 94–119, Jan. 2016.
[33] L. Rose, S. Lasaulce, S. Perlaza, and M. Debbah, "Learning equilibria with partial information in decentralized wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 136–142, Aug. 2011.
[34] O. Morgenstern and J. Von Neumann, *Theory of Games and Economic Behavior*. Princeton, NJ, USA: Princeton Univ. Press, 1953.
[35] R. J. Aumann, "Rationality and bounded rationality," in *Cooperation: Game-Theoretic Approaches*. Berlin, Germany: Springer, 1997, pp. 219–231.
[36] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.
[37] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," 2019, *arXiv:1911.10635*. [Online]. Available: http://arxiv.org/abs/1911.10635
[38] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, May 1996.
[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[40] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.
[41] V. Srivastava, J. O. Neel, A. B. MacKenzie, R. Menon, L. A. DaSilva, J. E. Hicks, J. H. Reed, and R. P. Gilles, "Using game theory to analyze wireless ad hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 7, nos. 1–4, pp. 46–56, 4th Quart., 2005.
[42] Y. Xu, A. Anpalagan, Q. Wu, L. Shen, Z. Gao, and J. Wang, "Decision-theoretic distributed channel selection for opportunistic spectrum access: Strategies, challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1689–1713, 4th Quart., 2013.
[43] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. Int. Conf. Comput. Commun. (INFOCOM)*, Turin, Italy, Apr. 2013, pp. 998–1006.
[44] A. Keshavarz-Haddad, E. Aryafar, M. Wang, and M. Chiang, "HetNets selection by clients: Convergence, efficiency, and practicality," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 406–419, Feb. 2017.
[45] Z. Du, B. Jiang, Q. Wu, Y. Xu, and K. Xu, *Towards User-Centric Intelligent Network Selection in 5G Heterogeneous Wireless Networks: A Reinforcement Learning Perspective*. Singapore: Springer, 2019.
[46] Z. Du, Q. Wu, P. Yang, Y. Xu, J. Wang, and Y.-D. Yao, "Exploiting user demand diversity in heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4142–4155, Aug. 2015.
[47] Z. Du, Q. Wu, P. Yang, Y. Xu, and Y.-D. Yao, "User-Demand-Aware wireless network selection: A localized cooperation approach," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4492–4507, Nov. 2014.
[48] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, p. 33275–33284, Jun. 2018.
[49] Q. Wu, Z. Du, P. Yang, Y.-D. Yao, and J. Wang, "Traffic-aware online network selection in heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 381–397, Jan. 2016.
[50] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 98, Madison, WI, USA: International Machine Learning Society, Jul. 1998, pp. 242–250.
[51] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[52] A. Nowé, P. Vrancx, and Y.-M. De Hauwere, "Game theory and multi-agent reinforcement learning," in *Reinforcement Learning*. Berlin, Germany: Springer, 2012, pp. 441–470.
[53] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel, "A unified game-theoretic approach to multiagent reinforcement learning," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 4193–4206.
[54] S. S. Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: An overview," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, vol. 2. London, U.K.: Springer, Sep. 2016, pp. 426–440.
[55] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
[56] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826–3839, Sep. 2020.
[57] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Auto. Agents Multi-Agent Syst.*, vol. 33, no. 6, pp. 750–797, Nov. 2019.
[58] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[59] H. Tembine, *Distributed Strategic Learning for Wireless Engineers*. Boca Raton, FL, USA: CRC Press, 2012.
[60] Z. Han, D. Niyato, W. Saad, and T. Başar, *Game Theory for Next Generation Wireless and Communication Networks: Modeling, Analysis, and Design*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[61] S. Lasaulce, M. Debbah, and E. Altman, "Methodologies for analyzing equilibria in wireless games," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 41–52, Sep. 2009.
[62] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," 2020, *arXiv:2011.03615*. [Online]. Available: http://arxiv.org/abs/2011.03615
[63] A. Stamou, N. Dimitriou, K. Kontovasilis, and S. Papavassiliou, "Autonomic handover management for heterogeneous networks in a future Internet context: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3274–3297, 4th Quart., 2019.
[64] M. Tayyab, X. Gelabert, and R. Jäntti, "A survey on handover management: From LTE to NR," *IEEE Access*, vol. 7, pp. 118907–118930, Aug. 2019.

[65] G. Liu, Y. Huang, Z. Chen, L. Liu, Q. Wang, and N. Li, "5G deployment: Standalone vs. non-standalone from the operator perspective," *IEEE Commun. Mag.*, vol. 58, no. 11, pp. 83–89, Nov. 2020.

[66] M. Simsek, M. Bennis, and I. Güvenç, "Context-aware mobility management in HetNets: A reinforcement learning approach," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, Mar. 2015, pp. 1536–1541.

[67] *E-UTRA Radio Resource Control (RRC); Protocol Specification (Release 9)*, document TS 36.331, 3GPP, 2009.

[68] *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-Configuring and Self-Optimizing Network (SON) Use Cases and Solutions*, document TR 36.902, 3GPP.

[69] *Evolved Universal Terrestrial Radio Access; Mobility Enhancements in Heterogeneous Networks*, document TR 36.839, 3GPP.

[70] V. Capdevielle, A. Feki, and E. Sorsy, "Joint interference management and handover optimization in LTE small cells network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Ottawa, ON, Canada, Jun. 2012, pp. 6769–6773.

[71] *Access to the 3GPP Evolved Packet Core (EPC) Via Non-3GPP Access Networks*, document 3GPP TS 24.302 v.10.4.0, 3GPP, 2011.

[72] D. Laselva, D. Lopez-Perez, M. Rinne, and T. Henttonen, "3GPP LTE-WLAN aggregation technologies: Functionalities and performance comparison," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 195–203, Mar. 2018.

[73] R. Bajracharya, R. Shrestha, R. Ali, A. Musaddiq, and S. W. Kim, "LWA in 5G: State-of-the-Art architecture, opportunities, and research challenges," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 134–141, Oct. 2018.

[74] *System Architecture for the 5G System*, document TS 23.501, v16.4, 3GPP.

[75] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, *TCP Extensions for Multipath Operation With Multiple Addresses*, document RFC 6824, 2013.

[76] J. R. Iyengar, P. D. Amer, and R. Stewart, "Concurrent multipath transfer using SCTP multihoming over independent end-to-end paths," *IEEE/ACM Trans. Netw.*, vol. 14, no. 5, pp. 951–964, Oct. 2006.

[77] Q. De Coninck and O. Bonaventure, "Multipath QUIC: Design and evaluation," in *Proc. Int. Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*. Seoul, South Korea, Nov. 2017, pp. 160–166.

[78] Y. Kang, C. Kim, D. An, and H. Yoon, "Multipath transmission control protocol–based multi-access traffic steering solution for 5G multimedia-centric network: Design and testbed system implementation," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 2, Feb. 2020, Art. no. 155014772090975.

[79] *Study on Access Traffic Steering, Switch and Splitting Support in the 5G System (5GS) Architecture*, TR 23.793, v1.1.0, 3GPP.

[80] P. C. Fishburn, "Utility theory for decision making," Res. Anal. Corp., McLean, VA, USA, Tech. Rep. AD0708563, 1970. [Online]. Available: https://apps.dtic.mil/sti/citations/AD0708563

[81] V. Rakocevic, J. Griffiths, and G. Cope, "Performance analysis of bandwidth allocation schemes in multiservice IP networks using utility functions," in *Teletraffic Science and Engineering*, vol. 4. Amsterdam, The Netherlands: Elsevier, Jan. 2001, pp. 233–243.

[82] F. Bari and V. C. Leung, "Use of non-monotonic utility in multi-attribute network selection," in *Wireless Technology*. Boston, MA, USA: Springer, Jul. 2009, pp. 21–39.

[83] Q.-T. Nguyen-Vuong, Y. Ghamri-Doudane, and N. Agoulmine, "On utility models for access network selection in wireless heterogeneous networks," in *Proc. NOMS-IEEE Netw. Operations Manage. Symp.*, Salvador, Brazil, Apr. 2008, pp. 144–151.

[84] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *IEEE Wireless Commun.*, vol. 11, no. 3, pp. 8–15, Jun. 2004.

[85] O. Ormond, J. Murphy, and G.-M. Muntean, "Utility-based intelligent network selection in beyond 3G systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 4, Istanbul, Turkey, Jun. 2006, pp. 1831–1836.

[86] F. Bari and V. Leung, "Automated network selection in a heterogeneous wireless network environment," *IEEE Netw.*, vol. 21, no. 1, pp. 34–40, Jan. 2007.

[87] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," *IEEE Wireless Commun.*, vol. 12, no. 3, pp. 42–48, Jun. 2005.

[88] E. Stevens-Navarro and V. W. Wong, "Comparison between vertical handoff decision algorithms for heterogeneous wireless networks," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, vol. 2. Melbourne, VIC, Australia, May 2006, pp. 947–951.

[89] F. Bari and V. Leung, "Multi-attribute network selection by iterative TOPSIS for heterogeneous wireless access," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2007, pp. 808–812.

[90] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Atlanta, GA, USA, Mar. 2004, pp. 653–658.

[91] F. Bari and V. Leung, "Application of ELECTRE to network selection in a hetereogeneous wireless network environment," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Hong Kong, Mar. 2007, pp. 3810–3815.

[92] M. Ma, A. Zhu, S. Guo, and Y. Yang, "Intelligent network selection algorithm for multi-service users in 5G heterogeneous network system: Nash Q-learning method," *IEEE Internet Things J.*, early access, p. 1, Apr. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9403383

[93] D. D. Nguyen, H. X. Nguyen, and L. B. White, "Reinforcement learning with network-assisted feedback for heterogeneous RAT selection," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6062–6076, Sep. 2017.

[94] L. Chen, "A distributed access point selection algorithm based on no-regret learning for wireless access networks," in *Proc. IEEE 71st Veh. Technol. Conf.*, Taipei, Taiwan, May 2010, pp. 1–5.

[95] K. Zhu, D. Niyato, and P. Wang, "Network selection in heterogeneous wireless networks: Evolution with incomplete information," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Sydney, SW, Australia, Apr. 2010, pp. 1–6.

[96] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 2008–2017, May 2009.

[97] Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover control in wireless systems via asynchronous multiuser deep reinforcement learning," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4296–4307, Dec. 2018.

[98] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Energy-efficient context-aware user association for outdoor small cell heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 1614–1619.

[99] D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Joint uplink and downlink user association for energy-efficient HetNets using Nash bargaining solution," in *Proc. IEEE 79th Veh. Technol. Conf. (VTC Spring)*, Seoul, South Korea, May 2014, pp. 1–5.

[100] H. Zhu, S. Wang, and D. Chen, "Energy-efficient user association for heterogenous cloud cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012, pp. 273–278.

[101] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3854–3868, Dec. 2016.

[102] I. Malanchini, M. Cesana, and N. Gatti, "Network selection and resource allocation games for wireless access networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 12, pp. 2427–2440, Dec. 2013.

[103] W. Liao, L. Wang, and J. Li, "Congestion game with inter-cell interference for cell selection in heterogeneous cellular network," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Shanghai, China, Jan. 2014, pp. 1–5.

[104] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[105] E. Liu, Q. Zhang, and K. K. Leung, "Asymptotic analysis of proportionally fair scheduling in Rayleigh fading," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1764–1775, Jun. 2011.

[106] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[107] L.-C. Tseng, F.-T. Chien, D. Zhang, R. Y. Chang, W.-H. Chung, and C. Huang, "Network selection in cognitive heterogeneous networks using stochastic learning," *IEEE Commun. Lett.*, vol. 17, no. 12, pp. 2304–2307, Dec. 2013.

[108] P. Naghavi, S. Hamed Rastegar, V. Shah-Mansouri, and H. Kebriaei, "Learning RAT selection game in 5G heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 52–55, Feb. 2016.

[109] X. Tan, X. Luan, Y. Cheng, A. Liu, and J. Wu, "Cell selection in two-tier femtocell networks using Q-learning algorithm," in *Proc. Int. Conf. Adv. Commun. Technol. (ICACT)*, Pyeongchang, South Korea, Feb. 2014, pp. 1031–1035.

[110] S. Maghsudi and E. Hossain, "Distributed user association in energy harvesting dense small cell networks: A mean-field multi-armed bandit approach," *IEEE Access*, vol. 5, pp. 3513–3523, Mar. 2017.

[111] S. Maghsudi and E. Hossain, "Distributed user association in energy harvesting small cell networks: A probabilistic bandit model," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1549–1563, Mar. 2017.
[112] M. Cesana, N. Gatti, and I. Malanchini, "Game theoretic analysis of wireless access network selection: Models, inefficiency bounds, and algorithms," in *Proc. 3rd Int. Conf. Perform. Eval. Methodologies Tools (ValueTools)*, Athens, Greece, Oct. 2008, pp. 1–10.
[113] R. Branzei, D. Dimitrov, and S. Tijs, *Models in Cooperative Game Theory*, vol. 556. Berlin, Germany: Springer, 2008.
[114] J. Nash, "Non-cooperative games," *Ann. Math.*, vol. 54, no. 2, pp. 286–295, Sep. 1951.
[115] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.
[116] S. Bayat, Y. Li, L. Song, and Z. Han, "Matching theory: Applications in wireless communications," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 103–122, Nov. 2016.
[117] Y. Narahari, *Game Theory and Mechanism Design*, vol. 4. Singapore: World Scientific, 2014.
[118] F. Pantisano, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Interference alignment for cooperative femtocell networks: A game-theoretic approach," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2233–2246, Nov. 2013.
[119] D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Performance evaluation of Nash bargaining solution based user association in HetNet," in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Lyon, France, Oct. 2013, pp. 571–577.
[120] W. Saad, Z. Han, R. Zheng, M. Debbah, and V. H. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Toronto, ON, Canada, May 2014, pp. 1096–1104.
[121] S. Bayat, R. H. Y. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027–3043, Aug. 2014.
[122] B. Ma, M. H. Cheung, V. W. S. Wong, and J. Huang, "Hybrid overlay/underlay cognitive femtocell networks: A game theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3259–3270, Jun. 2015.
[123] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wang, and T. Q. S. Quek, "Resource allocation for cognitive small cell networks: A cooperative bargaining game theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3481–3493, Jun. 2015.
[124] M. Anany, M. M. Elmesalawy, and A. M. Abd El-Haleem, "Matching game-based cell association in multi-RAT HetNet considering device requirements," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9774–9782, Dec. 2019.
[125] Y.-H. Yang, Y. Chen, C. Jiang, C.-Y. Wang, and K. J. R. Liu, "Wireless access network selection game with negative network externality," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5048–5060, Oct. 2013.
[126] J. C. Harsanyi, "Games with incomplete information played by 'Bayesian' players, I–III Part I. The basic model," *Manage. Sci.*, vol. 14, no. 3, pp. 159–182, Nov. 1967.
[127] J.-F. Mertens and S. Zamir, "Formulation of Bayesian analysis for games with incomplete information," *Int. J. Game Theory*, vol. 14, no. 1, pp. 1–29, Mar. 1985.
[128] S. Zamir, *Bayesian Games: Games with Incomplete Information*. New York, NY, USA: Springer, 2009.
[129] R. J. Aumann, "Correlated equilibrium as an expression of Bayesian rationality," *Econ., J. Econ. Soc.*, vol. 55, no. 1, pp. 1–18, Jan. 1987.
[130] D. Fudenberg and D. Levine, "Subgame-perfect equilibria of finite- and infinite-horizon games," *J. Econ. Theory*, vol. 31, no. 2, pp. 251–268, Dec. 1983.
[131] S. Cai, L. Duang, J. Wang, S. Zhou, and R. Zhang, "Incentive mechanism design for delayed WiFi offloading," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 3388–3393.
[132] X. Zhou, S. Feng, Z. Han, and Y. Liu, "Distributed user association and interference coordination in HetNets using stackelberg game," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 6431–6436.
[133] L. Zhong, M. Li, Y. Cao, and T. Jiang, "Stable user association and resource allocation based on stackelberg game in backhaul-constrained HetNets," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10239–10251, Oct. 2019.
[134] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *Int. J. Game Theory*, vol. 2, no. 1, pp. 65–67, Dec. 1973.
[135] D. Monderer and L. S. Shapley, "Potential games," *Games Econ. Behav.*, vol. 14, no. 1, pp. 124–143, May 1996.
[136] L. Wang, Y. Wang, Z. Ding, and X. Wang, "Cell selection game for densely-deployed sensor and mobile devices in 5G networks integrating heterogeneous cells and the Internet of Things," *Sensors*, vol. 15, no. 9, pp. 24230–24256, Sep. 2015.
[137] H. Ackermann, H. Röglin, and B. Vöcking, "Pure Nash equilibria in player-specific and weighted congestion games," *Theor. Comput. Sci.*, vol. 410, no. 17, pp. 1552–1563, Apr. 2009.
[138] T. Harks and M. Klimm, "On the existence of pure Nash equilibria in weighted congestion games," in *Proc. Int. Colloq. Automata, Lang., Program. (ICALP)*. Bordeaux, France: Springer, Jul. 2010, pp. 79–89.
[139] J. M. Smith, *Evolution Theory Games*. Cambridge, U.K.: Cambridge Univ. Press, 1982.
[140] J. M. Smith and G. R. Price, "The logic of animal conflict," *Nature*, vol. 246, no. 5427, p. 15, Nov. 1973.
[141] K. Tuyls, P. J. Hoen, and B. Vanschoenwinkel, "An evolutionary dynamical analysis of multi-agent learning in iterated games," *Auto. Agents Multi-Agent Syst.*, vol. 12, no. 1, pp. 115–153, Jan. 2006.
[142] P. D. Taylor and L. B. Jonker, "Evolutionary stable strategies and game dynamics," *Math. Biosci.*, vol. 40, nos. 1–2, pp. 145–156, Jul. 1978.
[143] A. M. Lyapunov, "The general problem of the stability of motion," *Int. J. Control*, vol. 55, no. 3, pp. 531–534, Mar. 1992.
[144] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, vol. 112. Springer, 2013.
[145] N. Sui, D. Zhang, W. Zhong, L. We, and Z. Zhang, "Evolutionary game theory based network selection for constrained heterogeneous networks," in *Proc. Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Shanghai, China, Apr. 2015, pp. 738–742.
[146] N. Sui, D. Zhang, W. Zhong, and C. Wang, "Network selection for heterogeneous wireless networks based on multiple attribute decision making and evolutionary game theory," in *Proc. 25th Wireless Opt. Commun. Conf. (WOCC)*, Chengdu, China, May 2016, pp. 1–5.
[147] A. De La Oliva, A. Banchs, I. Soto, T. Melia, and A. Vidal, "An overview of IEEE 802.21: Media-independent handover services," *IEEE Wireless Commun.*, vol. 15, no. 4, pp. 96–103, Aug. 2008.
[148] X. Wang, B. Liu, X. Su, X. Xu, and L. Xiao, "Evolutionary game based heterogeneous wireless network selection with multiple traffics in 5G," in *Proc. 26th Int. Conf. Telecommun. (ICT)*, Hanoi, Vietnam, Apr. 2019, pp. 80–84.
[149] S. Feng, D. Niyato, X. Lu, P. Wang, and D. I. Kim, "Dynamic model for network selection in next generation HetNets with memory-affecting rational users," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1365–1379, Apr. 2021.
[150] V. V. Tarasova and V. E. Tarasov, "Logistic map with memory from economic model," *Chaos, Solitons Fractals*, vol. 95, pp. 84–91, Feb. 2017.
[151] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci. USA*, vol. 39, no. 10, pp. 1095–1100, Oct. 1953.
[152] G. J. Laurent, L. Matignon, and N. Le Fort-Piat, "The world of independent learners is not Markovian," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 15, no. 1, pp. 55–64, Mar. 2011.
[153] T. Sandholm, "Perspectives on multiagent learning," *Artif. Intell.*, vol. 171, no. 7, pp. 382–391, May 2007.
[154] H. W. Kuhn and A. W. Tucker, *Contributions to Theory Games*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1953.
[155] A. M. Fink, "Equilibrium in a stochastic n-person game," *J. Sci. Hiroshima Univ., Ser. A-I (Math.)*, vol. 28, no. 1, pp. 89–93, 1964.
[156] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 1243–1254, Mar. 2008.
[157] P.-H. Tseng, K.-T. Feng, and C.-H. Huang, "POMDP-based cell selection schemes for wireless networks," *IEEE Commun. Lett.*, vol. 18, no. 5, pp. 797–800, May 2014.
[158] L. Gan, U. Topcu, and S. H. Low, "Optimal decentralized protocol for electric vehicle charging," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 940–951, May 2013.
[159] D. Fudenberg and D. M. Kreps, *Lectures on Learning and Equilibrium in Strategic Form Games*. Core Foundation, 1992.
[160] H. P. Young, "Learning by trial and error," *Games Econ. Behav.*, vol. 65, no. 2, pp. 626–643, Mar. 2009.
[161] R. Bellman, "Dynamic programming," *Science*, vol. 153, nos. 37–31, pp. 34–37, 1966.

[162] R. A. Howard, *Dynamic Programming and Markov Processes*. Hoboken, NJ, USA: Wiley, 1960.

[163] M. L. Puterman, *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, Aug. 2014.

[164] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.

[165] G. A. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," Dept. Eng., Univ. Cambridge, Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR 166, 1994, vol. 37. [Online]. Available: https://mailman.srv.cs.cmu.edu/pipermail/connectionists/1994-October/015828.html

[166] R. S. Sutton, "Generalization in reinforcement learning: Successful examples using sparse coarse coding," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Denver, CO, USA, Nov. 1995, pp. 1038–1044.

[167] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Comput.*, vol. 6, no. 6, pp. 703–710, Nov. 1994.

[168] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Mach. Learn.*, vol. 16, no. 3, pp. 185–202, Sep. 1994.

[169] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Mach. Learn.*, vol. 38, no. 3, pp. 287–308, Mar. 2000.

[170] S. Thrun, "The role of exploration in learning control," in *Handbook for Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. Florence, KY, USA: Van Nostrand Reinhold, Jun. 1992.

[171] C. Dhahri and T. Ohtsuki, "Q-learning cell selection for femto-cell networks: Single-and multi-user case," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. Anaheim, CA, USA, Dec. 2012, pp. 4975–4980.

[172] T. Kudo and T. Ohtsuki, "Cell selection using distributed Q-learning in heterogeneous networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Kaohsiung, Taiwan, Nov. 2013, pp. 1–6.

[173] J. S. Perez, S. K. Jayaweera, and S. Lane, "Machine learning aided cognitive RAT selection for 5G heterogeneous networks," in *Proc. IEEE Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, Istanbul, Turkey, Jun. 2017, pp. 1–5.

[174] E. Fakhfakh and S. Hamouda, "Optimised Q-learning for WiFi offloading in dense cellular networks," *IET Commun.*, vol. 11, no. 15, pp. 2380–2385, Aug. 2017.

[175] C. Wang, G. Wu, Z. Du, and B. Jiang, "Reinforcement learning based network selection for hybrid VLC and RF systems," in *Proc. MATEC Web Conf.*, vol. 173, 2018, p. 03014.

[176] X. Wang, J. Li, L. Wang, C. Yang, and Z. Han, "Intelligent user-centric network selection: A model-driven reinforcement learning framework," *IEEE Access*, vol. 7, pp. 21645–21661, Feb. 2019.

[177] B. Soleymani, A. Zamani, S. H. Rastegar, and V. Shah-Mansouri, "RAT selection based on association probability in 5G heterogeneous networks," in *Proc. IEEE Symp. Commun. Veh. Technol. (SCVT)*, Leuven, Belgium, Nov. 2017, pp. 1–6.

[178] X. Li, R. Cao, and J. Hao, "An adaptive learning based network selection approach for 5G dynamic environments," *Entropy*, vol. 20, no. 4, p. 236, Mar. 2018.

[179] Y. Xu, J. Chen, L. Ma, and G. Lang, "Q-learning based network selection for WCDMA/WLAN heterogeneous wireless networks," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Seoul, South Korea, May 2014, pp. 1–5.

[180] M. El Helou, M. Ibrahim, S. Lahoud, K. Khawam, D. Mezher, and B. Cousin, "A network-assisted approach for RAT selection in heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1055–1067, Jun. 2015.

[181] Y.-J. Liu, S.-M. Cheng, and Y.-L. Hsueh, "eNB selection for machine type communications using reinforcement learning based Markov decision process," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11330–11338, Dec. 2017.

[182] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[183] D. Lopez-Perez and X. Chu, "Inter-cell interference coordination for expanded region picocells in heterogeneous networks," in *Proc. 20th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Lahaina, HI, USA, Jul. 2011, pp. 1–6.

[184] I. Guvenc, M.-R. Jeong, I. Demirdogen, B. Kecicioglu, and F. Watanabe, "Range expansion and inter-cell interference coordination (ICIC) for picocell networks," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, Sep. 2011, pp. 1–6.

[185] *Hotspot 2.0 Technical Specification*, document TS v.1.0.0, WiFi Alliance, 2012.

[186] S. Buljore, H. Harada, S. Filin, P. Houze, K. Tsagkaris, O. Holland, K. Nolte, T. Farnham, and V. Ivanov, "Architecture and enablers for optimized radio resource usage in heterogeneous wireless access networks: The IEEE 1900.4 working group," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 122–129, Jan. 2009.

[187] M. Tokic, "Adaptive ε-greedy exploration in reinforcement learning based on value differences," in *Proc. German Conf. Adv. Artif. Intell. (KI)*. Karlsruhe, Germany: Springer, Sep. 2010, pp. 203–210.

[188] X. Wang, X. Su, and B. Liu, "A novel network selection approach in 5G heterogeneous networks using Q-Learning," in *Proc. 26th Int. Conf. Telecommun. (ICT)*, Hanoi, Vietnam, Apr. 2019, pp. 309–313.

[189] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, pp. 1039–1069, Nov. 2003.

[190] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multiagent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4539–4551, May 2018.

[191] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[192] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.

[193] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[194] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[195] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[196] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[197] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2015, *arXiv:1511.05952*. [Online]. Available: http://arxiv.org/abs/1511.05952

[198] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.* Phoenix, AZ, USA: Association for the Advancement of Artificial Intelligence, Feb. 2016, pp. 2094–2100.

[199] H. V. Hasselt, "Double Q-learning," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2010, pp. 2613–2621.

[200] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 48. New York, NY, USA: International Machine Learning Society, Jun. 2016, pp. 1995–2003.

[201] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 48. New York, NY, USA: International Machine Learning Society, Jun. 2016, pp. 1928–1937.

[202] C. Shen and M. van der Schaar, "A learning approach to frequent handover mitigations in 3GPP mobility protocols," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[203] C. Zhang, Z. Liu, B. Gu, K. Yamori, and Y. Tanaka, "A deep reinforcement learning based approach for cost- and energy-aware multi-flow mobile data offloading," *IEICE Trans. Commun.*, vol. E101.B, no. 7, pp. 1625–1634, Jul. 2018.

[204] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.
[205] Y. Fan and H. Li, "Distributed approximating global optimality with local reinforcement learning in HetNets," in *Proc. GLOBECOM IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–7.
[206] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.
[207] M. Sana, A. De Domenico, W. Yu, Y. Lostanlen, and E. C. Strinati, "Multi-agent reinforcement learning for adaptive user association in dynamic mmWave networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6520–6534, Oct. 2020.
[208] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Proc. AAAI Fall Symp. Sequential Decision Making Intell. Agents (AAAI-SDMIA)*, Arlington, VA, USA, Nov. 2015, pp. 29–37.
[209] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic Q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* San Diego, CA, USA, Nov. 2007, pp. 64–69.
[210] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, "Deep learning-based beam management and interference coordination in dense mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 592–603, Jan. 2019.
[211] Q. Zhang, Y.-C. Liang, and H. V. Poor, "Intelligent user association for symbiotic radio networks using deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4535–4548, Jul. 2020.
[212] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5G small cells," *IEEE Wireless Commun.*, vol. 23, no. 3, pp. 64–73, Jun. 2016.
[213] Y. Sun, G. Feng, S. Qin, Y.-C. Liang, and T.-S.-P. Yum, "The SMART handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1456–1468, Jun. 2018.
[214] S. Boldrini, L. De Nardis, G. Caso, M. Le, J. Fiorina, and M.-G. Di Benedetto, "MuMAB: A multi-armed bandit model for wireless network selection," *Algorithms*, vol. 11, no. 2, p. 13, Jan. 2018.
[215] S. Oh, A. M. Appavoo, and S. Gilbert, "Periodic bandits and wireless network selection," in *Proc. Int. Colloq. Automata, Lang., Program. (ICALP)*. Patras, Greece: European Association for Theoretical Computer Science, Jul. 2019.
[216] G. Nikolov, M. Kuhn, and B.-L. Wenning, "A contextual bandit approach to the interface selection problem," in *Proc. 24th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Zaragoza, Spain, Sep. 2019, pp. 1707–1714.
[217] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards," *Stoch. Syst.*, vol. 9, no. 4, pp. 319–337, Oct. 2019.
[218] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 41, no. 2, pp. 148–177, 1979.
[219] J. Chakravorty and A. Mahajan, "Multi-armed bandits, Gittins index, and its calculation," *Methods Appl. Statist. Clin. Trials Planning, Anal., Inferential Methods*, vol. 2, pp. 416–435, May 2014.
[220] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, no. A, pp. 287–298, 1988.
[221] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.
[222] L. Di Gregorio and V. Frascolla, "Handover optimality in heterogeneous networks," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, Dresden, Germany, Sep. 2019, pp. 365–370.
[223] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.
[224] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 235–256, 2002.
[225] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with Markovian rewards," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Sep. 2010, pp. 1675–1682.
[226] M. Huang, P. E. Caines, and R. P. Malhamé, "Large-population cost-coupled lqg problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-nash equilibria," *IEEE Trans. Autom. Control*, vol. 52, no. 9, pp. 1560–1571, Sep. 2007.
[227] R. Gummadi, R. Johari, and J. Y. Yu, "Mean field equilibria of multi armed bandit games," in *Proc. Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2012, p. 1110.
[228] R. Gummadi, R. Johari, S. Schmit, and J. Y. Yu. (Apr. 2013). *Mean Field Analysis of Multi-Armed Bandit Games*. [Online]. Available: https://ssrn.com/abstract=2045842, doi: 10.2139/ssrn.2045842.
[229] A. György, L. Kocsis, I. Szabó, and C. Szepesvári, "Continuous time associative bandit problems," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Hyderabad, India, Jan. 2007, pp. 830–835.
[230] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Proc. Eur. Conf. Mach. Learn. (ECML)*. Porto, Portugal: Springer, Oct. 2005, pp. 437–448.
[231] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Jan. 2002.
[232] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret bounds for sleeping experts and bandits," *Mach. Learn.*, vol. 80, nos. 2–3, pp. 245–272, Apr. 2010.
[233] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. Int. Conf. World Wide Web (WWW)*, Raleigh, NC, USA, Apr. 2010, pp. 661–670.
[234] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *J. Roy. Stat. Soc. C, Appl. Stat.*, vol. 41, no. 1, pp. 191–201, 1992.
[235] M. A. Khan, H. Tembine, and A. V. Vasilakos, "Game dynamics and cost of learning in heterogeneous 4G networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 198–213, Jan. 2012.
[236] P. Prabhavathi and L. Nithyanandan, "Network selection in wireless heterogeneous networks," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*. Melmaruvathur, India, Apr. 2013, pp. 357–361.
[237] A. Meetoo Appavoo, S. Gilbert, and K.-L. Tan, "Cooperation speeds surfing: Use co-bandit!," 2019, *arXiv:1901.07768*. [Online]. Available: http://arxiv.org/abs/1901.07768
[238] C. Tekin and M. Liu, "Performance and convergence of multi-user online learning," in *Proc. Int. Conf. Game Theory Netw. (GameNets)*. Shanghai, China: Springer, Apr. 2011, pp. 321–336.
[239] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2331–2345, Apr. 2014.
[240] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Algorithmic Learning Theory*, Lanzarote, Spain, Apr. 2018, pp. 56–92.
[241] I. Bistritz and A. Leshem, "Distributed multi-player bandits—A game of thrones approach," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 7222–7232.
[242] E. Boursier and V. Perchet, "SIC—MMAB: Synchronisation involves communication in multiplayer multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12071–12080.
[243] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, Sep. 2000.
[244] S. Hart and A. Mas-Colell, "A reinforcement procedure leading to correlated equilibrium," in *Economics Essays*. Springer, 2001, pp. 181–200.
[245] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 5, pp. 769–777, May 1994.
[246] R. R. Bush and F. Mosteller, *Stochastic Models for Learning*. Hoboken, NJ, USA: Wiley, 1955.
[247] H. Tembine, J.-Y. Le Boudec, R. El-Azouzi, and E. Altman, "Mean field asymptotics of Markov decision evolutionary games and teams," in *Proc. Int. Conf. Game Theory Netw. (GameNets)*. Istanbul, Turkey, May 2009, pp. 140–150.
[248] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *Proc. Annu. Symp. Theor. Aspects Comput. Sci. (STACS)*. Trier, Germany: Springer, Mar. 1999, pp. 404–413.
[249] E. Anshelevich, A. Dasgupta, J. Kleinberg, É. Tardos, T. Wexler, and T. Roughgarden, "The price of stability for network design with fair cost allocation," *SIAM J. Comput.*, vol. 38, no. 4, pp. 1602–1623, Jan. 2008.
[250] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Eastern Res. Lab., Digit. Equip. Corp., Hudson, MA, USA, Res. Rep. TR-301, 1984.
[251] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Tecnol.*, vol. 10, no. 2, p. 12, Jan. 2019.

[252] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.
[253] K. Hamidouche, A. T. Z. Kasgari, W. Saad, M. Bennis, and M. Debbah, "Collaborative artificial intelligence (AI) for user-cell association in ultra-dense cellular systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICCW)*, Kansas City, MI, USA, May 2018, pp. 1–6.
[254] P. Casgrain, B. Ning, and S. Jaimungal, "Deep Q-learning for Nash equilibria: Nash-DQN," 2019, *arXiv:1904.10554*. [Online]. Available: http://arxiv.org/abs/1904.10554
[255] Z. Fu, Z. Yang, Y. Chen, and Z. Wang, "Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games," 2019, *arXiv:1910.07498*. [Online]. Available: http://arxiv.org/abs/1910.07498
[256] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 6379–6390.
[257] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 32. Beijing, China: International Machine Learning Society, Jun. 2014, pp. 387–395.
[258] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/Jun. 2020.
[259] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J.-A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.
[260] P. Semasinghe, S. Maghsudi, and E. Hossain, "Game theoretic mechanisms for resource management in massive wireless IoT systems," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 121–127, Feb. 2017.
[261] X. Zhao, X. Li, Z. Xu, and T. Chen, "An optimal game approach for heterogeneous vehicular network selection with varying network performance," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 3, pp. 80–92, 2019.
[262] Q. Si, Z. Cheng, Y. Lin, L. Huang, and Y. Tang, "Network selection in heterogeneous vehicular network: A one-to-many matching approach," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, Antwerp, Belgium, May 2020, pp. 1–5.
[263] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, H. Won-Joo, and Z. Ding, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, Jun. 2020.
[264] J. Moura and D. Hutchison, "Game theory for multi-access edge computing: Survey, use cases, and future trends," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 260–288, 1st Quart., 2019.
[265] F. M. Abinader, E. P. L. Almeida, F. S. Chaves, A. M. Cavalcante, R. D. Vieira, R. C. D. Paiva, A. M. Sobrinho, S. Choudhury, E. Tuomaala, K. Doppler, and V. A. Sousa, "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 54–61, Nov. 2014.
[266] R. Zhang, M. Wang, L. X. Cai, Z. Zheng, X. Shen, and L.-L. Xie, "LTE-unlicensed: The future of spectrum aggregation for cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 150–159, Jun. 2015.
[267] S. Lagen, N. Patriciello, and L. Giupponi, "Cellular and Wi-Fi in unlicensed spectrum: Competition leading to convergence," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Levi, Finland, Mar. 2020, pp. 1–5.
[268] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "NR-U and IEEE 802.11 technologies coexistence in unlicensed mmWave spectrum: Models and evaluation," *IEEE Access*, vol. 8, pp. 71254–71271, Apr. 2020.
[269] C. Raiciu, C. Paasch, S. Barre, A. Ford, M. Honda, F. Duchene, O. Bonaventure, and M. Handley, "How hard can it Be? Designing and implementing a deployable multipath TCP," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, San Jose, CA, USA: USENIX, Apr. 2012, pp. 399–412.
[270] H. Wu, O. Alay, A. Brunstrom, S. Ferlin, and G. Caso, "Peekaboo: Learning-based multipath scheduling for dynamic heterogeneous environments," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2295–2310, Oct. 2020.
[271] C. Paasch, S. Ferlin, O. Alay, and O. Bonaventure, "Experimental evaluation of multipath TCP schedulers," in *Proc. ACM SIGCOMM, Workshop Capacity Sharing*. Chicago, IL, USA: ACM/SIGCOMM, Aug. 2014, pp. 27–32.
[272] H. Zhang, W. Li, S. Gao, X. Wang, and B. Ye, "ReLeS: A neural adaptive multipath scheduler based on deep reinforcement learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Paris, France, Apr. 2019, pp. 1648–1656.
[273] H. Wu, G. Caso, S. Ferlin, O. Alay, and A. Brunstrom, "Multipath scheduling for 5G networks: Evaluation and outlook," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 44–50, Apr. 2021.

GIUSEPPE CASO (Member, IEEE) received the B.Sc. degree in electrical engineering from the Federico II University of Naples, in 2009, and the M.Sc. and Ph.D. degrees from the Sapienza University of Rome, Italy, in 2012 and 2016, respectively. He has been a Postdoctoral Fellow with the Sapienza University of Rome, since 2018. He held various visiting positions with the Leibniz University of Hannover, King's College London, Technical University of Berlin, and Karlstad University. He is currently a Postdoctoral Fellow with the Department of Mobile Systems and Analytics, Simula Metropolitan, Oslo, Norway. His research interests include cognitive communications, HetNets, and the IoT technologies.

ÖZGÜ ALAY (Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from Middle East Technical University, Turkey, and the Ph.D. degree in electrical and computer engineering from the Tandon School of Engineering, New York University. She is currently an Associate Professor with the University of Oslo, Oslo, Norway, and the Head of the Department of Mobile Systems and Analytics, Simula Metropolitan, Oslo. She is the author of more than 70 peer-reviewed publications. Her research interests include 5G networks, multi-connectivity and multipath protocols, the IoT, drone communications, and multimedia systems.

GUIDO CARLO FERRANTE (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees *(summa cum laude)* in electrical engineering from the Sapienza Università di Roma, Italy, and the double Ph.D. degree in electrical engineering from the Sapienza Università di Roma, and CentraleSupélec, Gif-sur-Yvette, France, in 2015. He is a Senior Researcher with Ericsson Research, Stockholm, Sweden. Previously, he was a Postdoctoral Researcher with the Chalmers University of Technology, Gothenburg, Sweden; Massachusetts Institute of Technology (MIT), Cambridge, MA, USA; and Singapore University of Technology and Design (SUTD). He received the Italian National Telecommunications and Information Theory Group Award for Ph.D. thesis in the field of communication technologies (2015), and the SUTD-MIT Postdoctoral Fellowship (2015–2017).

LUCA DE NARDIS (Member, IEEE) received the Ph.D. degree from the Sapienza University of Rome, Italy, in 2005. In 2007, he was a Postdoctoral Fellow with the University of California, Berkeley. He is currently an Assistant Professor with the Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome. He has authored over 100 international peer-reviewed publications. His research interests focus on cognitive communications, medium access control, routing protocols, and positioning systems.

MARIA-GABRIELLA DI BENEDETTO (Fellow, IEEE) received the Ph.D. degree from the Sapienza University of Rome, Italy, in 1987. In 1991, she joined the Faculty of Engineering, Sapienza University of Rome, where she is currently a Full Professor of telecommunications. She held various visiting positions with the Massachusetts Institute of Technology, University of California, Berkeley, and the University of Paris XI. Her research interests include wireless communication systems, impulse radio communications, and speech. She is a fellow of the Radcliffe Institute for Advanced Study, Harvard University, Cambridge, MA, USA. In 1994, she received the Mac Kay Professorship Award from the University of California, Berkeley.

ANNA BRUNSTROM (Member, IEEE) received the B.Sc. degree in computer science and mathematics from Pepperdine University, CA, USA, in 1991, and the M.Sc. and Ph.D. degrees in computer science from the College of William & Mary, VA, USA, in 1993 and 1996, respectively. She joined the Department of Computer Science, Karlstad University, Sweden, in 1996, where she is currently a Full Professor and the Research Manager for the Distributed Systems and Communications Research Group. She has authored or coauthored over 170 international peer-reviewed publications. Her research interests include Internet architectures and protocols, multipath and low-latency communications, and performance evaluation of mobile broadband systems, including 5G.