# Game Dynamics and Cost of Learning in Heterogeneous 4G Networks

Manzoor Ahmed Khan, Hamidou Tembine and Athanasios V. Vasilakos

***Abstract*—In this paper, we study game dynamics and learning schemes for heterogeneous 4G networks. We introduce a novel learning scheme called *cost-to-learn* that incorporates the cost to switch, the switching delay, and the cost of changing to a new action and, captures the realistic behavior of the users that we have experimented on OPNET simulations. Considering a dynamic and uncertain environment where the users and operators have only a numerical value of their own payoffs as information, we construct various heterogeneous combined fully distributed payoff and strategy reinforcement learning (CODIPAS-RL): the users try to learn their own optimal payoff and their optimal strategy simultaneously. We establish the asymptotic pseudo-trajectories as solution of differential equations. Using evolutionary game dynamics, we prove the convergence and stability properties in specific classes of dynamic robust games. We provide various numerical examples and OPNET simulations in the context network selection in wireless local area networks (WLAN) and Long Term Evolution (LTE).**

***Index Terms*—Game dynamics, strategic learning, heterogeneous 4G networks, cost of learning.**

## I. Introduction

ONE OF THE REASONS to consider dynamic scenarios in evolving networks is that they seem to show up in reality more often. Network traffic, routing, congestion games, security games have been applied to networks that involve few or large number of selfish users such as Internet routing, peer-to-peer file sharing systems, etc. However, in most of the studies a static network model is considered which includes a game which is framed over static network, static user demand and a fixed iterative learning scheme. As the complexity of the existing system grows, and the environment cannot be assumed to be constant, we need to study and explore the dynamic behavior of such systems which involve not only the time dependencies and the state of the environment but also the variability of the demands, the uncertainty of the system parameters, the random activity of the users, the time delays, error and noise in the measurement over long-run interactions, etc.

In many dynamic interactions, one would like to have a learning and adaptive procedure that does not require any information about the other users' actions or payoffs and as little memory (small number of parameters in term of past own-actions and past own-payoffs) as possible. Such a rule is said to be *uncoupled* or *fully distributed*. However, it has been shown in [1] that for a large class of games, no such general algorithm causes the users' period-by-period behavior to converge to Nash equilibrium (no user can improve its payoff by unilateral deviation). Hence, most of the time, there is no guarantee that the behaviors of fully distributed learning algorithms and dynamics will come close to Nash equilibrium. By introducing public signals (but irrelevant-payoff signals) into the interaction, each user (player) can choose his/her action according to her observation of the value of the signal. Then, a strategy assigns an action to every possible observation user can make. If no user would want to deviate from the recommended strategy (assuming the others don't deviate), the distribution is called a correlated equilibrium. The works in [2], [3] showed that *regret-minimizing procedures* can cause the empirical frequency distribution of play to converge to the set of correlated equilibria. Note that the set of correlated equilibria is convex and includes the convex hull of the set of Nash equilibria.

### A. Game dynamics

Many networking and communication games are subject to uncertainty (i.e., robust games). Uncertainties may come from the measurements, the noisy observations, the computation errors or the incomplete information. In robust games with a large number of actions, users are inherently faced with limitations in both their observational and computational capabilities. Accordingly, users in such games need to make their decisions using algorithms that accommodate limitations in information gathering and processing. This disqualifies some of the well known decision making models (such as *fictitious play*, best reply, gradient descent, model-based algorithms etc) in which each user must monitor the actions of every other user and must optimize over a high dimensional probability space (cartesian product of the action spaces). The authors in [4] proposed a modified version of the fictitious play called joint-action fictitious play with inertia and proved its convergence in potential games and network congestion games using the finite improvement path (FIP) property. Note that in the finite improvement path procedure only one user moves at a given time slot (simultaneous moves are not allowed). For this reason, the FIP is not adapted if the network does not follow a prescribed rule evolution with observation capabilities. One of the well-known learning schemes for simultaneous-move games is the *interactive trial and error learning*. In [5], it is shown that the interactive trial and error learning, implements

Manuscript received 15 January 2011; revised 15 July 2011. This work has been started when the first author was visiting Supelec, Ecole Supérieure d'Electricité, France.

M. A. Khan is with Technische Universität Berlin, Germany (e-mail: manzoor-ahmed.khan@dai-labor.de).

H. Tembine is with Ecole Supérieure d'Electricité, Supelec, France (e-mail: tembine@ieee.org).

A. V. Vasilakos with University of Western Macedonia, Greece (e-mail: vasilako@ath.forthnet.gr).

Digital Object Identifier 10.1109/JSAC.2012.120118.

Nash equilibrium behavior in any game with generic payoffs and which has at least one pure Nash equilibrium. The interactive trial and error learning is a completely uncoupled learning rule, such that, when used by all users in a game, period-by-period play comes close to pure Nash equilibrium play a high proportion of the time, provided that the game has such an equilibrium and the payoffs satisfy an interdependency condition. However, in games without pure Nash equilibrium (such as matching pennies, penalty games, many security games etc.), the interactive trial and error learning does not implement Nash equilibria. Another point is that even if the trial-and-error process is at a pure Nash equilibrium, it can move from this point and the process restarts again.

Since we know from the *Nash theorem* that any finite game in strategic form has at least one equilibrium in mixed strategies and the same result can be applied to finite robust games under suitable condition on the mathematical expectation, it remains a question of algorithms for computing one of them and the selection of the most efficient equilibrium (if any). In the line of mixed equilibria search (including pure equilibria), several stochastic learning procedures have been proposed. Strategy reinforcement learning and dynamics in finite games have been studied for both pure and mixed equilibria. Most of these works used stochastic approximation techniques [6], [7], [8], [9] to derive ordinary differential equations (ODE) equivalent to the adjusted replicator dynamics [10]. By studying the orbits of the replicator dynamics, one can get some convergence/divergence and stability/instability properties of the system. However the replicator dynamics may not lead to *approximate equilibria* even in simple games. Convergence properties in special class of games such as weakly acyclic games and best-response potential games can be found in [11]. Recently, distributed learning algorithms and feedback based update rules have been extensively developed in networking and communication systems. Closely related works on network selection and dynamics can be found in [12], [13], [14]. The authors in [13] focus on service provider selection, where users' service provider selection criteria encamps the subscription fee and coverage. Authors model the competition between operators using game theoretic approach and study the impact of user types distribution within the coverage area in fixed & dynamic configurations. [14] studies user subscription dynamics, revenue maximization, and equilibrium characteristics in two different markets (i.e., monopoly and duopoly). Although the research works [13], [14] discuss the co-existence of network technologies, however they do not discuss the technical realization of the technologies integration. We, on the other hand have provided and extensively implemented the technical solution i.e., IMS functional entities in the core network, integration of LTE & WLAN network technologies based on 3GPP standards, and IPv6 based mobility management etc. The consequence of such extensive technical solution implementation is the increased confidence level in attained results specifically if the network selection model involves dynamic wireless parameters. It should be noted that in the referred research literature, user evaluation of the network selection is based on abstract functions i.e., not concretely taking the technical QoS indices into account. However, we use user utility function, that captures user satisfaction with respect to both technical and economical aspects. We also validate the proposed user satisfaction against the objective measurements for three different types of applications i.e., Voice over IP (VoIP), Video and File Transfer Protocol (FTP). It should be noted that the objective measurements were carried out in the extensively developed measurement setup following ITU-T and 3GPP standards.

Delayed evolutionary game dynamics have been studied in [15], [16], [17], [18] but in continuous time. The authors have shown that an *evolutionary stable strategy* (which is robust to invasions by small fraction of users) can be unstable for large time delays and they provided sufficient conditions of stability of delayed Aloha-like systems.

Different from *distributed learning optimization*, we use the term *strategic learning* [19]. By strategic learning, we mean how users are able to learn about the dynamic environment under their complex and interdependent strategies - the convergence of learning of each user depends on the others and so on.

### B. Case of interest of this paper

In this paper, we focus on hybrid and combined strategic learning for general-sum stochastic dynamic games with incomplete information and action-independent state transition with the following novelties:

- In contrast to the standard learning approaches widely studied in the literature where the users follow the same predetermined scheme, here we relax this assumption and the users do not need to follow the same learning patterns. We propose different learning schemes that the users can adopt. This leads to *heterogeneous learning*. Our motivation for heterogeneous learning follows from the observation that, in heterogeneous wireless systems, the users may not see the environment in the same way, they may have different capabilities and different adaptation degrees. Thus, it is important to take into consideration these differences when analyzing the behavior of the wireless system. The heterogeneity is crucial in terms of convergence of certain systems.
- Each user does not need to update his strategy at each iteration. The updating times are random and unknown by the users. Usually, in the iterative learning schemes the time slots during which the user updates are fixed. Here we do not restrict to fixed updating time. This is because some users come in or exit temporarily, and it may be costly to update or for some other reasons, the users may prefer to update their strategies at another time. One may think that if some of the user does not update often, the strategic learning process will be slower in terms of convergence time; this statement is less clear because the off-line users may indirectly help the online users to converge and, when they wake-up they respond to an already converged system, and so on.
- Each user can be in active mode or in sleep mode. When a user is active, he/she can select from a set of learning patterns to update his strategies and/or estimations. The user can change their learning pattern during the interaction. This leads to a *hybrid learning*.

TABLE I
SUMMARY OF NOTATIONS

| Symbol | Meaning |
|---|---|
| $\mathbb{R}^k$ | $k-$dimensional Euclidean space |
| $\mathcal{W} \subseteq \mathbb{R}^k$ | state space |
| $\mathcal{N}$ | set of potential users (finite or infinite) |
| $\mathcal{B}^n(t)$ | random set of active users at time $t$. |
| $\mathcal{A}_j$ | set of actions of user $j$ |
| $s_j \in \mathcal{A}_j$ | a generic element of $\mathcal{A}_j$ |
| $\mathcal{X}_j := \Delta(\mathcal{A}_j)$ | set of probability distributions over $\mathcal{A}_j$ |
| $a_{j,t} \in \mathcal{A}_j$ | action of the user $j$ at time $t$ |
| $\mathbf{x}_{j,t} \in \mathcal{X}_j$ | strategy of the user $j$ at $t$ |
| $u_{j,t}$ | perceived payoff by user $j$ at $t$ |
| $\hat{\mathbf{u}}_{j,t} \in \mathbb{R}^{\lvert\mathcal{A}_j\rvert}$ | estimated payoff vector of user $j$ at $t$ |
| $l^2$ | space of sequences $\{\lambda_t\}_{t\geq 0}$, $\sum_{t\in\mathbb{N}} \lvert\lambda_t\rvert^2 < +\infty$ |
| $l^1$ | space of sequences $\{\lambda_t\}_{t\geq 0}$, $\sum_{t\in\mathbb{N}} \lvert\lambda_t\rvert < +\infty$ |
| $(\lambda_{j,t}, \nu_{j,t})$ | learning rates of user $j$ at $t$ |
| $m_t^p(.)$ | Mean field limit at time $t$ |

- We propose a *cost of learning* CODIPAS-RL which takes into consideration the cost of moves from one action to another one. In the context of technology selection, the cost of learning is very important, it can represent the delay needed to change a technology or a production or an upgrade cost.
- We establish a connection between the asymptotic pseudo-trajectory of the learning schemes to the *hybrid evolutionary game dynamics* developed in [20].
- In contrast to the standard learning frameworks developed in the literature which are limited to a finite and fixed number of users, we extend our methodology to large systems with multiple classes of populations. This allows us to address the "curse of dimensionality" problems when the size of the interacting system is very large. Finally, different *mean field learning* are proposed using *Fokker-Planck-Kolmogorov* equations. The case of noisy and time delayed payoffs is also discussed.
- Our theoretical findings are illustrated numerically in heterogeneous wireless networks with multiple classes of users and multiple technologies: wireless local area networks (WLAN) and long term evolution (LTE) using Mathematica and OPNET Simulation.

### C. Structure

The paper is structured as follows. In Section 2 we describe the model of non-zero-sum dynamic game, and present different learning patterns. Then, we develop hybrid and delayed learning schemes in noisy and dynamic environment. Section 3 presents mean field learning. Section 4 focuses on learning under noisy strategy. In section 5, we apply our learning framework in heterogeneous wireless networks. Section 6 concludes the paper. The proofs are given in Appendix.

We summarize some of the notations in Table I.

## II. THE SETTING

### A. Description of the dynamic environment

We examine a system with a finite number of *potential users*. The set of users is denoted by $\mathcal{N} = \{1, 2, \ldots, n\}$, $n = |\mathcal{N}|$. The number $n$ can be 10, $10^4$ or $10^6$. Each user has a finite number of actions denoted by $\mathcal{A}_j$ (which can be arbitrary large). Time is discrete and the space of time is $\mathbb{N} = \{0, 1, 2, \ldots\}$. A user does not necessarily interact at all the time steps. Each user can be in one of the two modes: *active mode* or *sleep mode*. The set of users interacting at the current time is the set of active users $\mathcal{B}^n(t) \subseteq \mathcal{N}$. This time-varying set is unknown to the users. When user is in active mode, he/she does an experiment, and gets a measurement or a reaction to his decision, denoted $u_{j,t} \in \mathbb{R}$ (this may be delayed as we will see). Let $\mathcal{X}_j := \Delta(\mathcal{A}_j)$ be the set of probability distributions over $\mathcal{A}_j$ i.e the simplex of $\mathbb{R}^{|\mathcal{A}_j|}$. The number $u_{j,t} \in \mathbb{R}$ is the realization of a random variable $\tilde{U}_{j,t}$ which depends on the state of nature $\mathbf{w}_t \in \mathcal{W}$ and the action of the users where the set $\mathcal{W}$ is a subset of a finite dimensional Euclidean space. Each *active user* $j$ updates her/his current strategy $\mathbf{x}_{j,t+1} \in \Delta(\mathcal{A}_j)$ based on his experiment and its prediction for his future interaction via the payoff estimation $\hat{\mathbf{u}}_{j,t+1} \in \mathbb{R}^{|\mathcal{A}_j|}$.

This leads into the class of dynamic games with unknown payoff function and with imperfect monitoring (the last decisions of the other users are not observed). A payoff in the long-run interaction is the average payoff which we assume to have a limit. In that case, under the stationary strategies, the limiting of the average payoff can be expressed as an expected game i.e the game with payoff $v_j : \prod_{j'\in\mathcal{N}} \mathcal{X}_{j'} \longrightarrow \mathbb{R}$,

$$v_j(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n} \left(\mathbb{E}\tilde{U}_j\right)$$

**Assumptions on user's information:** The only information assumed is that each user is able to observe or to measure a noisy value of its payoff when she/he is active and update its strategy based on this measurement.

Note that the users do not need to know their own action space in advance. Each user can learn his action space (using for example exploration techniques). In that case, we need to add an exploration phase or a progressive exploration during the dynamic game. The result is that if the all the actions have been *explored* and sufficiently *exploited* and if the learning rate are well-chosen then the prediction can be "good" enough.

Next, we describe how the dynamic robust game evolves.

### B. Description of the dynamic game

The dynamic robust game is described as follows:

- At time slot $t = 0$, $\mathcal{B}^n(0)$ is the set of active users. The set $\mathcal{B}^n(0)$ is not known by the users. We assume that each user has its internal state in $\{0, 1\}$. The number 1 corresponds to the case where $j \in \mathcal{B}^n(0)$, and 0 otherwise. Each user $j \in \mathcal{B}^n(0)$ chooses an action $a_{j,0} \in \mathcal{A}_j$. The set $\mathcal{A}_j$ is not assumed to be known in advance by user $j$, we assume that he can explore progressively during the interactions. He measures a numerical noisy value of its payoff which corresponds to a realization of the random variables depending on the actions of the other users and the state of the nature etc. He initializes its estimation to $\hat{\mathbf{u}}_{j,0}$. The non-active users get zero.
- At time slot $t$, each user $j \in \mathcal{B}^n(t)$ has an estimation of his payoffs, chooses an action based its own-experiences and experiments a new strategy. Each user $j$ measures/observes an output $u_{j,t} \in \mathbb{R}$, (eventually

after some time delay). Based on this target $u_{j,t}$, the user $j$ updates its estimation vector $\hat{\mathbf{u}}_{j,t} \in \mathbb{R}^{|\mathcal{A}_j|}$ and built a strategy $\mathbf{x}_{j,t+1} \in \mathcal{X}_j$ for his next interaction. The strategy at $t+1$, $\mathbf{x}_{j,t+1}$ is a function only of $\mathbf{x}_{j,t}, \hat{\mathbf{u}}_{j,t}$ and the most recent target value. Since the users do not interact always, each user has its own clock which counts the activity of that user. At time step $t$, the clock of user $j$ is $\theta_j(t) = \sum_{t' \leq t} \mathbb{1}_{\{j \in \mathcal{B}^n(t')\}}$. We assume $\liminf_{t \longrightarrow \infty} \theta_j(t)/t > 0$.

Note that the exact value of the state of the nature at time $t$ and the previous strategies $\mathbf{x}_{-j,t-1} := (\mathbf{x}_{k,t-1})_{k \neq j}$ of the other users and their past payoffs $\mathbf{u}_{-j,t-1} := (u_{k,t-1})_{k \neq j}$ are unknown to user $j$ at time $t$.

- The game moves to $t+1$.

In addition, we extend the framework to the delayed payoff measurement case. This means that, the perceived payoffs at time $t$ are not the instantaneous payoff but the noisy value of the payoff at $t - \tau_j$ i.e $u_{j,t-\tau_j}$.

In order to define rigourously the dynamic robust game, we need some preliminaries. Next, we introduce the notions of histories, strategies and payoffs (performance metrics). The payoff is associated to a (behavioral) strategy profile which is a collection of mapping from the set of histories to the available actions at the current time.

**Histories** A user's information consists of his (own) past activities, own-actions and measured own-payoffs. A private history up to $t$ for user $j$ is a collection

$$h_{j,t} = (b_{j,0}, a_{j,0}, u_{j,0}, b_{j,1}, a_{j,1}, u_{j,1}, \ldots, b_{j,t-1}, a_{j,t-1}, u_{j,t-1})$$

in the set $H_{j,t} := (\{0,1\} \times \mathcal{A}_j \times \mathbb{R})^t$. where $b_{j,t} = \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}}$ which is 1 if $j$ is active at time $t$ and 0 otherwise.

**Behavioral Strategy** A behavioral strategy for user $j$ is a mapping $\tilde{\tau}_j : \bigcup_{t \geq 0} H_{j,t} \longrightarrow \mathcal{X}_j$. We denote by $\Sigma_j$ the set of behavioral strategies of user $j$.

The set of complete histories of the dynamic robust game after $t$ stages is $H_t = (2^{\mathcal{N}} \times \mathcal{W} \times \prod_{j \in \mathcal{N}} \mathcal{A}_j \times \mathbb{R}^n)^t$, it describes the set of active users, the states, the chosen actions and the received payoffs for all the users at all past stages before $t$. The set $2^{\mathcal{N}}$ denotes the set of all the subsets of $\mathcal{N}$ (except the empty set). A behavioral strategy profile $\tilde{\tau} = (\tilde{\tau}_j)_{j \in \mathcal{N}} \in \prod_j \Sigma_j$ and a initial state $\mathbf{w}$ induce a probability distribution $P_{\mathbf{w},\tilde{\tau}}$ on the set of plays $H_\infty = (\mathcal{W} \times \prod_j \mathcal{A}_j \times \mathbb{R}^n)^{\mathbb{N}}$.

**Payoffs** Assume that $\mathbf{w}, \mathcal{B}^n$ are independent and independent of the strategy profiles. For a given $\mathbf{w}, \mathcal{B}^n$, we denote

$$U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{x}) := \mathbb{E}_{(\mathbf{x}_k)_{k \in \mathcal{B}^n}} \tilde{U}_j^{\mathcal{B}^n}(\mathbf{w}, (a_k)_{k \in \mathcal{B}^n}).$$

Let $\mathbb{E}_{\mathbf{w},\mathcal{B}^n}$ be the mathematical expectation relatively to the measure generated by the random variables $\mathbf{w}, \mathcal{B}^n$. Then, the expected payoff can be written as $\mathbb{E}_{\mathbf{w},\mathcal{B}^n} \tilde{U}_j^{\mathcal{B}^n}(.,.)$.

We focus on the limiting of the average payoff i.e $F_{j,T} = \frac{1}{T} \sum_{t=1}^{T} u_{j,t}$. The long-term payoff reduces to

$$\frac{1}{\sum_{t=1}^{T} \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}}} \sum_{t=1}^{T} u_{j,t} \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}},$$

when considering only the activity of user $j$. We assume that we do not have short-term users or equivalently the probability for a user $j$ to be active is strictly positive. Given a initial state $\mathbf{w}$ and a strategy profile $\tilde{\tau}$, the payoff of user $j$ is the superior limiting of the Cesaro-mean payoff $\mathbb{E}_{\mathbf{w},\tilde{\tau},\mathcal{B}^n} F_{j,T}$. We assume that $\mathbb{E}_{\mathbf{w},\tilde{\tau},\mathcal{B}^n} F_{j,T}$ has a limit. Then, the expected payoff of an active user $j$ is denoted by $v_j(e_{s_j}, \mathbf{x}_{-j}) = \mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, e_{s_j}, \mathbf{x}_{-j})$ where $e_{s_j}$ is the vector unit with 1 at the position of $s_j$ and zero otherwise.

**Definition 1** (Expected robust game)**.** *We define the expected robust game as* $\left(\mathcal{N}, (\mathcal{X}_j)_{j \in \mathcal{N}}, \mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, .)\right)$.

**Definition 2.** *A strategy profile* $(\mathbf{x}_j)_{j \in \mathcal{N}} \in \prod_{j=1}^n \mathcal{X}_j$ *is a (mixed) state-independent equilibrium for the expected robust game if and only if* $\forall j \in \mathcal{N},\ \forall \mathbf{y}_j \in \mathcal{X}_j$,

$$\mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{y}_j, \mathbf{x}_{-j}) \leq \mathbb{E}_{\mathbf{w},\mathcal{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{x}_j, \mathbf{x}_{-j}), \qquad (1)$$

The existence of solution of Equation (1) is equivalent to the existence of solution of the following *variational inequality problem*: find $\mathbf{x}$ such that

$$\langle \mathbf{x} - \mathbf{y}, V(\mathbf{x}) \rangle \geq 0,\ \forall \mathbf{y} \in \prod_j \mathcal{X}_j$$

where $\langle ., . \rangle$ is the inner product, $V(\mathbf{x}) = [V_1(\mathbf{x}), \ldots, V_n(\mathbf{x})]$,

$$V_j(\mathbf{x}) = [\mathbb{E}_{\mathbf{w},\mathcal{B}} U_j^{\mathcal{B}}(\mathbf{w}, e_{s_j}, \mathbf{x}_{-j})]_{s_j \in \mathcal{A}_j}.$$

**Remark 1.** *Note that an equilibrium of the expected robust game may not be an equilibrium (of the robust game) at each time slot. This is because* $\mathbf{x}$ *is an equilibrium for expected robust game does not imply that* $\mathbf{x}$ *is an equilibrium of the game* $\mathcal{G}(\mathbf{w})$ *for some state* $\mathbf{w}$ *and the set of active users may vary.*

**Lemma 1.** *Assume that* $\mathcal{W}$ *is compact. Then, The expected robust game with unknown state and variable number of interacting users has at least one (state-independent) equilibrium.*

The existence of such equilibrium points is guaranteed since the mappings $v_j : (\mathbf{x}_j, \mathbf{x}_{-j}) \longmapsto \mathbb{E}_{\mathbf{w},\mathcal{B}} U_j^{\mathcal{B}}(\mathbf{w}, \mathbf{x}_j, \mathbf{x}_{-j})$ is jointly continuous, quasi-concave in $\mathbf{x}_j$, the spaces $\mathcal{X}_j$, are non-empty, convex and compact. Then, the result follows by using Kakutani fixed-point theorem or by applying Nash theorem to the expected robust game [21].

Since we have existence of state-independent equilibrium under suitable conditions, we seek for heterogeneous and combined algorithms to locate the equilibria.

## III. CODIPAS-RL WITH RANDOM UPDATES

We propose a delayed hybrid COmbined fully DIstributed PAyoff and Strategy Reinforcement Learning in the following form: (hybrid-delayed-CODIPAS-RL) (See equation at the top of the following page) where $\hat{\mathbf{u}}_{j,t} = (\hat{u}_{j,t}(s_j))_{s_j \in \mathcal{A}_j} \in \mathbb{R}^{|\mathcal{A}_j|}$ is a vector payoff estimation of user $j$ at time $t$. Note that when user $j$ uses $a_{j,t} = s_j$, he observes only his measurement corresponding to that action but not those of the other actions $s'_j \neq s_j$. Hence he needs to estimate/predict them via the vector $\hat{\mathbf{u}}_{j,t+1}$. The functions $K^1$ and $\lambda$ are based on estimated payoffs and perceived measured payoff (delayed and noisy) such that the invariance of simplex is preserved almost surely. The function $K_j^1$ defines the strategy learning pattern of user $j$ and $\lambda_{j,\theta_j(t)}$ is its strategy learning rate. If at least two of the functions $K_j$ are different then we refer to *heterogeneous*

$$\left\{\begin{array}{c}\mathbf{x}_{j,t+1}(s_j)-\mathbf{x}_{j,t}(s_j)=\\ \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}}\sum_{l\in\mathcal{L}}\mathbb{1}_{\{l_{j,t}=l\}}K^{1,(l)}_{j,s_j}(\lambda_{j,\theta_j(t)},a_{j,t},u_{j,t-\tau_j},\hat{\mathbf{u}}_{j,t},\mathbf{x}_{j,t}),\\ \hat{\mathbf{u}}_{j,t+1}(s_j)-\hat{\mathbf{u}}_{j,t}(s_j)=\\ \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}}K^2_{j,s_j}(\nu_{j,\theta_j(t)},a_{j,t},u_{j,t-\tau_j},\hat{\mathbf{u}}_{j,t},\mathbf{x}_{j,t}),\\ j\in\mathcal{N},t\geq 0,a_{j,t}\in\mathcal{A}_j,s_j\in\mathcal{A}_j,\\ \theta_j(t+1)=\theta_j(t)+\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}},\\ t\geq 0,\ \mathcal{B}^n(t)\subseteq\mathcal{N},\\ \mathbf{x}_{j,0}\in\mathcal{X}_j,\hat{\mathbf{u}}_{j,0}\in\mathbb{R}^{|\mathcal{A}_j|}.\end{array}\right.$$

*learning* in the sense that the learning schemes of the users are different. If all the $K_j^1$ are identical but the learning rates $\lambda_j$ are different, we refer to *learning with different speed*: slow learners, medium or fast learners. Note that the term $\lambda_{j,\theta_j(t)}$ is used instead of $\lambda_{j,t}$ because the global clock $[t]$ is not known by user $j$ (he knows only how many times he has been active, the activity of others is not known by $j$). $\theta_j(t)$ is a random variable that determines the local clock of $j$. Thus, the updates are asynchronous. The functions $K_j^2$, and $\nu_j$ are well-chosen in order to have a good estimation of the payoffs. $\tau_j$ is a time delay associated to user $j$ in its payoff measurement. The payoff $u_{j,t-\tau_j}$ at $t-\tau_j$ is perceived at time $t$. We examine the case where the users can choose different CODIPAS-RL patterns during the dynamic game. They can select among a set of CODIPAS-RLs denoted by $\mathcal{L}_1,\ldots,\mathcal{L}_m, m\geq 1$. The resulting learning scheme is called *hybrid CODIPAS-RL*. The term $l_{j,t}$ is the CODIPAS-RL pattern chosen by user $j$ at time $t$.

### A. CODIPAS-RL patterns with random updates

In order to examine the above dynamic game we provide below some examples of learning patterns in which each user learns according to a specific CODIPAS-RL scheme.

*1) Bush-Mosteller based CODIPAS-RL: $\mathcal{L}_1$:* The learning pattern $\mathcal{L}_1$ is given by

$$x_{j,t+1}(s_j)-x_{j,t}(s_j)=\lambda_{\theta_j(t)}\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}}\times \frac{u_{j,t}-\Gamma_j}{\sup_{\mathbf{a},w}|U_j(w,a)-\Gamma_j|}\left(\mathbb{1}_{\{a_{j,t}=s_j\}}-\mathbf{x}_{j,t}(s_j)\right), \quad (2)$$

$$\hat{u}_{j,t+1}(s_j)-\hat{u}_{j,t}(s_j)= \nu_{\theta_j(t)}\mathbb{1}_{\{a_{j,t}=s_j,j\in\mathcal{B}^n(t)\}}\left(u_{j,t}-\hat{u}_{j,t}(s_j)\right) \quad (3)$$

$$\theta_j(t+1)=\theta_j(t)+\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \quad (4)$$

where $\Gamma_j$ is a reference level of $j$. The first equation of $\mathcal{L}_1$ is widely studied in machine learning and have been initially proposed by Bush & Mosteller in 1949-55 [22]. The second equation of $\mathcal{L}_1$ is a payoff estimation for the experimented action by the users. Combined together one gets a specific combined fully distributed payoff and strategy reinforcement learning based on Bush-Mosteller reinforcement learning.

*2) Boltzmann-Gibbs based CODIPAS-RL: $\mathcal{L}_2$:*

$$x_{j,t+1}(s_j)-x_{j,t}(s_j)=\lambda_{\theta_j(t)}\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}}\times \left(\frac{e^{\frac{1}{\epsilon_j}\hat{\mathbf{u}}_{j,t}(s_j)}}{\sum_{s'_j\in\mathcal{A}_j}e^{\frac{1}{\epsilon_j}\hat{\mathbf{u}}_{j,t}(s'_j)}}-x_{j,t}(s_j)\right), \quad (5)$$

$$\hat{u}_{j,t+1}(s_j)-\hat{u}_{j,t}(s_j)= \nu_{\theta_j(t)}\mathbb{1}_{\{a_{j,t}=s_j,j\in\mathcal{B}^n(t)\}}\left(u_{j,t}-\hat{u}_{j,t}(s_j)\right) \quad (6)$$

$$\theta_j(t+1)=\theta_j(t)+\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \quad (7)$$

The strategy learning (5) of $\mathcal{L}_2$ is a Boltzmann-Gibbs based reinforcement learning. Note that the Boltzmann-Gibbs distribution can be obtained from the maximization of the perturbed payoff $U_j+\epsilon_j H_j$ where $H_j$ is the entropy function i.e., $H_j(\mathbf{x}_j)=-\sum_{s_j\in\mathcal{A}_j}x_j(s_j)\ln x_j(s_j)$. It is a smooth best response function. Here the Boltzmann-Gibbs mapping is based on the payoff estimation (the exact payoff vector is not known, only one component of a noisy value is observed). We denote the Boltzmann-Gibbs strategy by

$$\tilde{\beta}_{j,\epsilon_j}(\hat{u}_{j,t})(s_j)=\frac{e^{\frac{1}{\epsilon_j}\hat{\mathbf{u}}_{j,t}(s_j)}}{\sum_{s'_j\in\mathcal{A}_j}e^{\frac{1}{\epsilon_j}\hat{\mathbf{u}}_{j,t}(s'_j)}}$$

and the smooth best response to $\mathbf{x}_{-j,t}$ (also called Logit rule, Gibbs sampling or Glauber dynamics) is given by

$$\beta_{j,\epsilon_j}(\mathbf{x}_{-j,t})(s_j)=\frac{e^{\frac{1}{\epsilon_j}v_j(\mathbf{e}_{s_j},\mathbf{x}_{-j,t})}}{\sum_{s'_j\in\mathcal{A}_j}e^{\frac{1}{\epsilon_j}v_j(\mathbf{e}_{s'_j},\mathbf{x}_{-j,t})}}.$$

*3) Imitative BG CODIPAS-RL: $\mathcal{L}_3$:*

$$x_{j,t+1}(s_j)-x_{j,t}(s_j)=\lambda_{\theta_j(t)}\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}}x_{j,t}(s_j)\times \left(\frac{e^{\frac{1}{\epsilon_j}\hat{\mathbf{u}}_{j,t}(s_j)}}{\sum_{s'_j\in\mathcal{A}_j}x_{j,t}(s'_j)e^{\frac{1}{\epsilon_j}\hat{\mathbf{u}}_{j,t}(s'_j)}}-1\right), \quad (8)$$

$$\hat{u}_{j,t+1}(s_j)-\hat{u}_{j,t}(s_j)= \nu_{\theta_j(t)}\mathbb{1}_{\{a_{j,t}=s_j,j\in\mathcal{B}^n(t)\}}\left(u_{j,t}-\hat{u}_{j,t}(s_j)\right) \quad (9)$$

$$\theta_j(t+1)=\theta_j(t)+\mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \quad (10)$$

The strategy learning (8) of $\mathcal{L}_3$ is an imitative Boltzmann-Gibbs based reinforcement learning. The imitation here consists to play an action with a probability proportional to the previous uses of that action. The imitation learning leads to an *imitative evolutionary game dynamics*.

*4) Multiplicative Weighted Imitative CODIPAS-RL: $\mathcal{L}_4$:*

$$x_{j,t+1}(s_j) - x_{j,t}(s_j) = \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} x_{j,t}(s_j) \times \left( \frac{(1+\lambda_{\theta_j(t)})^{\hat{\mathbf{u}}_{j,t}(s_j)}}{\sum_{s'_j\in\mathcal{A}_j} x_{j,t}(s'_j)(1+\lambda_{\theta_j(t)})^{\hat{\mathbf{u}}_{j,t}(s'_j)}} - 1 \right), \quad (11)$$

$$\hat{u}_{j,t+1}(s_j) - \hat{u}_{j,t}(s_j) = \nu_{\theta_j(t)} \mathbb{1}_{\{a_{j,t}=s_j, j\in\mathcal{B}^n(t)\}} \left(u_{j,t} - \hat{u}_{j,t}(s_j)\right) \quad (12)$$

$$\theta_j(t+1) = \theta_j(t) + \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \quad (13)$$

The strategy learning (11) of $\mathcal{L}_4$ is a learning rate weighted imitative reinforcement learning. The main difference with $\mathcal{L}_2$ and $\mathcal{L}_3$ is that there is no parameter $\epsilon_j$. The interior outcomes are necessarily exact equilibria of the expected (not approximated equilibria as in $\mathcal{L}_2$). It is easy to show that [23] this leads to replicator dynamics (thus its relative interior stationary points are Nash equilibria).

*5) Weakened fictitious play based CODIAPS-RL: $\mathcal{L}_5$:*

$$x_{j,t+1}(s_j) - x_{j,t}(s_j) \in \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \times \left( (1-\epsilon_t)\delta_{\arg\max_{s'_j} \hat{u}_{j,t}(s'_j)} + \epsilon_t \frac{\mathbb{1}}{|\mathcal{A}_j|} - x_{j,t}(s_j) \right), \quad (14)$$

$$\hat{u}_{j,t+1}(s_j) - \hat{u}_{j,t}(s_j) = \nu_{\theta_j(t)} \mathbb{1}_{\{a_{j,t}=s_j, j\in\mathcal{B}^n(t)\}} \left(u_{j,t} - \hat{u}_{j,t}(s_j)\right) \quad (15)$$

$$\theta_j(t+1) = \theta_j(t) + \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \quad (16)$$

The last learning pattern $\mathcal{L}_5$ is a combined learning based on the weakened fictitious play with asynchronous clocks. Here a user does not observe the action played by the other at the previous step and the payoff function is not known. Each user estimates its payoff function via the equations (15). The equation (14) consists to play one of the action with the best estimation $\hat{u}_{j,t}$ with probability $(1-\epsilon_t)$ and plays an arbitrary action with probability $\epsilon_t$.

*6) Payoff Learning:* We mention some payoff learning based the idea of CODIPAS-RL: • $\mathcal{PL}_1$ **No-regret based CODIPAS-RL**

$$x_{j,t+1}(s_j) - x_{j,t}(s_j) = \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} R_t(s_j), \quad (17)$$

$$\hat{u}_{j,t+1}(s_j) - \hat{u}_{j,t}(s_j) = \nu_{\theta_j(t)} \mathbb{1}_{\{a_{j,t}=s_j, j\in\mathcal{B}^n(t)\}} \left(u_{j,t} - \hat{u}_{j,t}(s_j)\right) \quad (18)$$

$$\theta_j(t+1) = \theta_j(t) + \mathbb{1}_{\{j\in\mathcal{B}^n(t)\}} \quad (19)$$

$$R_t(s_j) = \frac{\phi([\hat{u}_{j,t}(s_j) - u_{j,t}]_+)}{\sum_{s'_j} \phi([\hat{u}_{j,t}(s'_j) - u_{j,t}]_+)} \quad (20)$$

Here the function $\phi$ is a positive function defined in $\mathbb{R}$. The frequency of plays of strategy learning based on non-regret rule is known to be convergent to the set of correlated equilibria [1]. Here the non-regret is based on the estimations.

• $\mathcal{PL}_2$ : **Imitative No-regret based CODIPAS-RL**

See Eq. 21-24 next page.

### *B. Main results*

We introduce the following assumptions. $[H2]$, $\forall j \in \mathcal{N}$, $\liminf_{t\longrightarrow\infty} \frac{\theta_j(t)}{t} > 0$

$[H3]$ $\lambda_t \geq 0$, $\lambda \in l^2\backslash l^1$, $\mathbb{E}\left(M_{j,t+1} \mid \mathcal{F}_t\right) = 0$, $\forall\, j$, $\mathbb{E}\left(\| M_{j,t+1} \|^2\right) \leq c_1 \left[1 + \sup_{t'\leq t} \| \mathbf{x}_{t'} \|^2\right]$ where $c_1 > 0$ is a constant.

It is important to mention that these assumptions H2-H3 are standard assumptions in stochastic approximations for almost sure convergence. However the vanishing learning rate can be time-consuming. In order to design fast convergent learning algorithms, *constant* learning rate ($\lambda_t = \lambda$) can be used as well, and convergence in law can be proved under suitable conditions. In this case the expectation of the gap between the solution of differential equations and the stochastic process is in order of the constant learning rate i.e $O(\lambda)$. In particular, if $\lambda \longrightarrow 0$ one has a weak convergence. Below we give the main results for time-varying learning rate under H2-H3.

**Proposition 1** (proportional rates)**.** *Suppose H2-H3 and consider proportional learning rates (the ratio is relatively similar and non-vanishing). Then, The asymptotic pseudo-trajectory of the hybrid-delayed-CODIPAS-RL is given by*

$$\begin{cases} \frac{d}{dt}\mathbf{x}_{j,t}(s_j) = g_{j,t} \sum_{l\in\mathcal{L}} p_{j,t,l} f^{(l)}_{j,s_j}(\mathbf{x}_{j,t}, \hat{\mathbf{u}}_{j,t}), \\ \frac{d}{dt}\hat{\mathbf{u}}_{j,t}(s_j) = \bar{g}_{j,t}\left(\mathbb{E}_{\mathbf{w},\mathbb{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{e}_{j,s_j}, \mathbf{x}_{-j,t} - \hat{\mathbf{u}}_{j,t}(s_j))\right) \\ t \geq 0 \\ \mathbf{x}_{j,0} \in \mathcal{X}_j, \hat{\mathbf{u}}_{j,0} \in \mathbb{R}^{|\mathcal{A}_j|}. \end{cases}$$

*where $g_{j,t}$ is the limiting of the expected value of $\frac{\lambda_{j,t}}{\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t},\mu_{j',t})} \mathbb{1}_{j\in\mathcal{B}^n(t)}$. The function $\bar{g}_{j,t}$ is the limiting of the expected value of $\frac{\mu_{j,t}}{\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t},\mu_{j',t})} \mathbb{1}_{j\in\mathcal{B}^n(t)}$. The function $f^{(l)}_j$ the expected value of $K^{1,(l)}_j$ when $\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t}, \mu_{j',t})$ goes to zero. $p_{j,t,l}$ is the probability of the event $\{l_{j,t} = l\}$.*

**Consequence for wireless networking games** The Proposition 1 says that under suitable conditions of the learning rate, the above learning schemes can be studied by their differential equation counterparts, and the result applies directly to autonomous self-organizing networks with randomly changing channel states, variable number of interacting users and random updating time slots. The next result establishes heterogeneous learning convergence and capture the impact of different behavior of the users.

**Proposition 2** (heterogenous rates)**.** *Assume (i) H2-H3 and Assume that the payoff-learning rates are faster than strategy learning rates i.e* $[H4]$ $\lambda_t \geq 0, \nu_t \geq 0$, $(\lambda, \nu) \in (l^2\backslash l^1)^2$, $\frac{\lambda_t}{\nu_t} \longrightarrow 0$. *(ii) the payoff-learning converges globally to a unique point for any intermediary permutation of variables of the players. Then, hybrid-delayed-CODIPAS-RL scheme with variable number of players has the asymptotic pseudo trajectory of the following non-autonomous system:*

$$\begin{cases} \dot{x}_{j,t}(s_j) = g_{j,t} \sum_{l\in\mathcal{L}} p_{j,t,l} f^{(l)}_{j,s_j}(\mathbf{x}_{j,t}, \mathbb{E}_{\mathbf{w},\mathcal{B}} U_j^{\mathcal{B}}(\mathbf{w}, ., \mathbf{x}_{-j,t})) \\ x_j(s_j) > 0 \Longrightarrow \hat{\mathbf{u}}_{j,t}(s_j) \longrightarrow \mathbb{E}_{\mathbf{w},\mathcal{B}} U_j^{\mathcal{B}}(\mathbf{w}, \mathbf{e}_{s_j}, \mathbf{x}_{-j}) \end{cases}$$

We define two properties:

• NS: Nash stationary property refers to the configuration in which the set of Nash equilibria of the expected game coincide with the rest points (stationary points) of the resulting hybrid dynamics.

• PC: Positive Correlation property refers to the configuration where the covariance between the strategies generated by the dynamics and the payoff is positive. i.e $F(x) \neq 0 \Longrightarrow \sum_{j,s_j} u_j(e_{s_j}, \mathbf{x}_{-j}) F_{j,s_j}(\mathbf{x}) > 0$ where $F$ is the drift of the dynamics. We say that the expected robust game is a

$$x_{j,t+1}(s_j) - x_{j,t}(s_j) = \lambda_{\theta_j(t)} \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}} \left( IR_t(s_j) - x_{j,t}(s_j) \right), \tag{21}$$

$$\hat{u}_{j,t+1}(s_j) - \hat{u}_{j,t}(s_j) = \nu_{\theta_j(t)} \mathbb{1}_{\{a_{j,t}=s_j, j \in \mathcal{B}^n(t)\}} \left( u_{j,t} - \hat{u}_{j,t}(s_j) \right) \tag{22}$$

$$\theta_j(t+1) = \theta_j(t) + \mathbb{1}_{\{j \in \mathcal{B}^n(t)\}} \tag{23}$$

$$IR_t(s_j) := \frac{x_{j,t}(s_j)\phi([\hat{u}_{j,t}(s_j) - u_{j,t}]_+)}{\sum_{s'_j} x_{j,t}(s'_j)\phi([\hat{u}_{j,t}(s'_j) - u_{j,t}]_+)} \tag{24}$$

potential game if there exists a regular function $W$ such that $u_j(e_{s_j}, \mathbf{x}_{-j}) = \frac{\partial}{\partial x_j(s_j)} W(\mathbf{x})$.

**Proposition 3.**
- *(i) If the homogeneous learning are all NSs. Then the heterogeneous learning satisfy (NS)*
- *(ii) If the homogeneous are all (PC) then the heterogeneous are too. (example: Replicator and Smith dynamics). If the potential function serves as Lyapunov in all these dynamics then global convergence holds for the heterogeneous learning.*
- *(iii) The heterogeneous time-scaling leads to a new class of dynamics obtained by composition.*
- *The result of (i) and (ii) extends to hybrid learning (at each active time, the player can select among a set of learning patterns).*
- *(iv) Consider a hybrid of (PCs). If the support of the hybrid learning contains at least one (NS) then the "non-Nash rest points" are eliminated.*
- *(v) the result (iv) extends to evolutionary games.*

**Impact of these results in wireless networking games** Many networking and wireless communications are dynamic in nature and the number of users in the system are randomly changing due users mobility, channel variation, weather conditions, technologies and protocols evolutions etc. In many cases, the games have specific structures such as *aggregative games, pseudo-potential games, supermodular games*. This result gives the convergence of heterogeneous learning to equilibria in dynamic robust potential games but also in dynamic monotone games. Note that these two classes of games include many topology-based network congestion games, network selection games, frequency selection, concave routing games etc.

## IV. Mean field hybrid learning

The standard learning schemes are limited to the finite and fixed number of players case. As a consequence, the resulting differential equations leads to high dimensional system when the size of network is large [24]. In this subsection we show how to extend the learning framework to large number of players called *mean field learning*.

### A. Learning under noisy strategy

Following the above lines, one can generalize the CODIPAS-RL in the context of Itô's stochastic differential equation (SDE). Typically, the case where the strategy learning has the following form: $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda_t(f(\mathbf{x}_t, \hat{\mathbf{u}}_t) + M_{t+1}) + \sqrt{\lambda_t}\sigma(\mathbf{x}_t, \hat{\mathbf{u}}_t)\xi_t$, where $\xi_t$ is a difference of independent Brownian, can be seen as an Euler scheme of the Itô's SDE:

$$d\mathbf{x}_{j,t} = f_j(\mathbf{x}_t, \hat{\mathbf{u}}_t)dt + \sigma_j(\mathbf{x}_t, \hat{\mathbf{u}}_t)d\mathbb{B}_{j,t}, \tag{25}$$

where $\mathbb{B}_{j,t}$ is a standard Brownian motion in $\mathbb{R}^{|\mathcal{A}_j|}$. This leads stochastic evolutionary game dynamics where the stochastic stability of equilibria can be used to find robustness of the system under stochastic fluctuations. Note that the distribution of the noisy strategy-learning (25) or equivalently the mean field learning can be characterized by a solution of the following partial differential equation called Fokker-Planck-Kolmogorov equation

$$\partial_t m_{j,t}(x) + div(f_j m_{j,t}) - \frac{1}{2}\partial^2_{xx}(\sigma_j \sigma_j^t m_{j,t}) = 0. \tag{26}$$

where $div$ is the divergence operator and $\partial^2_{xx}$ is the second derivative operator with the respect to $x$. Particular case of this class of dynamics are *evolutionary game dynamics with diffusion terms*. We refer to [23] for the derivation of these equations which require the theory of distribution and integration by parts.

### B. Cost of learning CODIPAS-RL

In this subsection we introduce a novel way of learning under switching cost called *Cost-To-Learn CODIPAS-RL*. Usually in learning in games or in machine learning (reinforcement learning, best reply, fictitious play, gradient-descent/ascent based learning, nonmodel gradient estimation, Q-learning etc), the cost of switching between the actions, the cost of experimenting with another option are not taken into consideration. In this section we take these issues into account and study their effects in the learning outcome. The idea is that *it can be very costly to learn quickly and learning can take some time*. When a player changes its action, there is cost for that. In our scenario, the learning cost can arise in three different situations: (i) handover switch, (ii) codec-switchover, (iii) joint handover-and-codec switch-over. In a more general setting, one can think about a cost to have a new technology or a cost to produce a specific product. The reason for this cost of learning approach is that, in many situations, changing, improving the performance, the quality of experience of a user, guaranteeing to a quality of service etc has cost. At a given time $t$, if user $j$ changed its selection (codec, handover etc) i.e if user $j$ moves, its objective function is translated from the standard utility plus an additional cost for moving from the old configuration to the new one. Then, there is no additional cost to learn if the action remains the same.

## V. Application to Heterogeneous Wireless Networks

### A. User-centric network selection

It is envisioned that in future mobile communication paradigm, the decision of network selection will be delegated to the users to exploit the best available characteristics of different network technologies and network providers, with the objective of increased satisfaction. The consequence of such user-centric network selection paradigm is users' short term contractual agreements with the operators. These contracts will basically be driven by satisfaction level of users with operator services. In order to more accurately express the user satisfaction, the term Quality of Service (QoS) has been extended to include more subjective and also application specific measures beyond traditional technical parameters, giving rise to the Quality of Experience (QoE) concept. Intuitively this provides the representation and modeling of user satisfaction function, which captures user satisfaction for both technical (delay, jitter, packetloss, throughput, etc.) and economical (service cost etc.) aspects. It should be noted that in broader sense user preferences over different technical and economical aspects can be translated into user QoE. This motivates the authors to categorize the users into three categories namely *Excellent*, *Good*, and *Fair* users. We define these user types on the basis of the user preferences over different involved parameters for network selection decision making. For instance an excellent user is motivated to pay higher service prices for an excellent service quality and does not care much for service prices. One can think of putting business users in this category. On the other hand a fair user prefers cheaper services and remains ready to compromise on service quality, an example of such user may be a student user. On the similar lines a good user stands midway between the two mentioned user types. When it comes to the differentiation of users on the basis of service quality, we mean the user perceived application QoS or QoE. Thus to differentiate users on these lines for both real and non-real time traffic types, we need different bounded regions of QoE values e.g., Mean Opinion Score (MOS) values (ranges between$[0-5]$, with $zero$ representing the most irritated user and $5$ representing the most satisfied user) are the numeric values capturing the user QoE for Voice over IP (VoIP) applications. We generalize this QoE metric to all the traffic types and set the MOS value bounds for different user types on the similar lines as Modified E-model sets its R-factor values to distribute users in very satisfied, satisfied, and some users dissatisfied etc. categories. In this work the MOS values $4.3$ and above, $3.59 \sim 4.3$ and $3.1 \sim 3.59$ represent the excellent, good, and fair users respectively. One may object the suitability of MOS values as QoE metric for non-real-time applications e.g., TCP based FTP traffic, and can argue throughput or delivery response time to be the suitable QoE measurement metric for such traffic types. In this case a transformation or scaling function may be used to scale the user satisfaction to the MOS value range i.e., $[0-5]$.

It should be noted that MOS value is the function of QoS measurement metric delay, jitter, and packet loss. However we focus that user QoE is the function of both technical and economical parameters. In this connection, we have suggested analytical satisfaction function [25], which takes into account both the mentioned aspects. We validate the QoE prediction of user satisfaction function against the objective measurements (typically for technical parameters [1]). Whereas the user satisfaction for the service cost is captured through the following function. $u_k(\pi_k^c) = \tilde{\mu}_k^c - \frac{\tilde{\mu}_k^c}{1-e^{\tilde{\pi}_k^c}} e^{-\tilde{\pi}_k^c \epsilon}$, where $\tilde{\mu}_k^c$ represents the maximum satisfaction level of user type $k$ for service type $c$ , and $\tilde{\pi}_k^c$ is the private valuation of service by user, and $\epsilon$ represents the price sensitivity of user. We have developed and extensively validated the utility based user satisfaction model against subjective (from experiments) and objective (using network simulator) for different dynamics of the wireless environment. The validation results showed that the proposed user satisfaction function predicts the user QoE with the correlation $0.923$, the details of the user satisfaction function modeling is out of the scope of this paper. However we summarize in Table II the ranges of technical parameter values attained from user satisfaction function and validated against the subjective and objective measurement results. The range for service costs for different types of user follow the pattern $\pi_{excellent} > \pi_{good} > \pi_{fair}$ and the corresponding user satisfaction from the offered price is computed by the pricing function mentioned in earlier.

TABLE II
QoS parameters and ranges from the user utility function

| | G 7.11 Codec | (96kbps) | |
|---|---|---|---|
| Parameters | Range | MOS | Category |
| Delay | $0ms \sim 50ms$ | 4.3 and above | Excellent |
| | 50ms $\sim$ 200ms | $3.59 \sim 4.3$ | Good |
| | 200ms $\sim$ 300ms | $3.1 \sim 3.59$ | Fair |
| Packet Loss | $0\% \sim 3\%$ | 4.3 and above | Excellent |
| | $3\% \sim 10\%$ | $3.59 \sim 4.3$ | Good |
| | $10\% \sim 18\%$ | $3.1 \sim 3.59$ | Fair |
| | Non-real-time | FTP | |
| Delay | $0ms \sim 40ms$ | 4.3 and above | Excellent |
| | $40ms \sim 50ms$ | $3.59 \sim 4.3$ | Good |
| | $50ms \sim 60ms$ | $3.1 \sim 3.59$ | Fair |
| Packet Loss | $0\% \sim 3\%$ | 4.3 and above | Excellent |
| | $3\% \sim 3.5\%$ | $3.59 \sim 4.3$ | Good |
| | $3.4\% \sim 4\%$ | $3.1 \sim 3.59$ | Fair |
| | Video | (x264) | |
| Delay | $0ms \sim 20ms$ | 4.3 and above | Excellent |
| | $20ms \sim 60ms$ | $3.59 \sim 4.3$ | Good |
| | $60ms \sim 90ms$ | $3.1 \sim 3.59$ | Fair |
| Packet Loss | $0\% \sim 0.4\%$ | 4.3 and above | Excellent |
| | $0.4\% \sim 1.5\%$ | $3.59 \sim 4.3$ | Good |
| | $1.5\% \sim 3.5\%$ | $3.1 \sim 3.59$ | Fair |

*1) Proposed Architecture for 4G user-centric paradigm:* In this subsection, we briefly highlight the possible architectural issues associated with the implementation of proposed user-centric approach and we also propose the architecture and explain its functional components and the integration of architectural components.

Given the basic assumption of users having no long-term contractual agreements with the operators, the natural questions one can think of are:

- When the user mobile is turned on, what will be her default connection operator?

[1]As using modified E-model, PESQ etc. for VoIP application one can capture user QoE for delay, packet loss, and jitter parameters, therefore validation could be carried out for these parameters. Similarly for video applications we use PSNR and different codecs for validation.

- Assuming there exists a default connection operator, how and where in the technical infrastructure is the network selection decision executed?
- Who is responsible for user authentication in the system?
- How does an operator integrated the 3GPP and non-3GPP (trusted and untrusted) technologies providing host based network selection facility?

To address the highlighted issues, the proposed IP Multi-media Subsystem (IMS) based architecture should meet the following requirements:

- It should support the involvement of third party and extension of services from different IMS core operators.
- It should delegate the service subscription control to the end-users. Hence the operators many implement any Session Initiation Protocol (SIP) services by researching the end user demands. The users should have freedom to subscribe to any service given that it could be delivered using IMS control plane.
- It should enable dynamic partnership of operators with the third party.
- Owing to business contract of User Equipment (UE) with the third party, the complete user profile should be maintained in the Home Subscriber Server (HSS) of third party and each operator should only receive the service specific user data from HSS.

we suggest a model similar to the semi-wall garden business model, where a network provider acts as a bitpipe plus a service broker. This model is open to all parties and its service panoply is as rich as the internet and is a converged business model. We propose IMS (IP Multimedia Subsystems) based on SIP and other IETF protocols to realize the proposed user-centric approach. We assume that there exists a neutral and trusted third party, Telecommunication Service Provider (TSP), the TSP has no Radio Access Network (RAN) infrastructure, however it contains few functional components of IMS architecture namely Proxy Call Session Control Function (P-CSCF), Serving Call Session Control Function (S-CSCF), Interrogating Call Session Control Function (I-CSCF), Application server (AS), and HSS[26]. As discussed in the OPNET simulation settings section that operator communication footprint in any geographical area comprises of heterogeneous wireless access technologies(3GPP and non-3GPP). Figure 1 details the integration of operator RANs to the operator core network and operator's integration to the trusted third party operator.

**Sequence of Actions** Users send the service requests to the third party, who then transmits the requests to the available operators. Operators submit the service offers including QoS indices values and service costs. Third party on behalf of users suggests the best available network(s) to the users for requested service. Third party takes care service billing.

**Architecture functional entities and their interaction** We now briefly discuss functional entities and their interaction with each other.

*Trusted third party functional entity:* This entity is a basically a SIP application server, which processes SIP messages formulated using SIP MESSAGE method from UE and operators. In the proposed configuration, the SIP application on $3^{rd}$ party is enabled to understand XML (Extensible Markup Language) messages, which are enclosed in the message body of the proposed SIP MESSAGE method. An example illustrating such MESSAGE for user is shown in Figure 2. After the third party receives this message from UE through UE $\rightarrow$ default operator IMS core network $\rightarrow$ third party (I,S)-CSCF $\rightarrow$ third party AS, the registration process is initiated and completed. The third party AS then extracts user service request for body of SIP message from UE and triggers the network selection decision mechanism. The consequence of computation at decision maker, the third party generates one of the two possible responses. i.e., i) the network selection algorithm has successfully resulted in resource allocation and service price decision. These decisions are executed by generating two simultaneous responses; of which one goes to UE indicating the successful operation and the other is sent to the operator.

*Operator functional entity:* Integration of operator technologies: Owing to the maturity of current communication paradigm, it is needless to highlight the importance of heterogeneous wireless technologies and their co-existence to extend service to end-consumers. When it comes to heterogeneous wireless technologies, one can discern various prevailing standards in the current communication market, such as 3GPP, non-3GPP, 3GPP2 etc. The current communication market is framed to accept the integrated 3GPP and non-3GPP technologies. We follow the 3GPP standard for such integration as shown in Figure 1. We now consider the following use cases of the proposed architectural solution; basically non-3GPP technologies can be integrated with 3GPP technologies through one of the three interfaces $(S2a, S2b, S2c)$ provided by EPC/SAE(Evolved Packet Core / System Architecture Evolution). The description of each of the interface is as follows: i) S2a - it provides the integration path between trusted non-3GPP IP networks and 3GPP networks. In this case the mobility is handled by the network based mobility solution e.g., Proxy MIPv6. ii) S2b - It provides the integration path between untrusted non-3GPP IP networks and 3GPP networks. In this case mobility is handled by network based mobility solution. iii) S2c - It provides the integration between both trusted and un-trusted non-3GPP IP networks and 3GPP networks. In this case the mobility is handled by the host based mobility solution e.g., Dual stack MIPv6.

**Interaction of operator with third party** In case the operator wants to participate in the game, it must configure the Operator functional entity to register its parameters for indicated time length. This entity can formulate one SIP message for one service or multiple cost and offered quality information for multiple services using one SIP message, however in this case care should be taken that SIP message size does not exceed the upper bound defined by IETF standards for SIP messages. operator functional entity should maintain a record of all its sent messages because such information can not be retrieved from the third party application server. This entity, when declared as the chosen operator receives a notification from the third party as a SIP MESSAGE containing the service type, user preferences, user identity etc. It should be noted that in response to SIP MESSAGE from this entity, a SIP specified acknowledgement response must be sent by the third party that

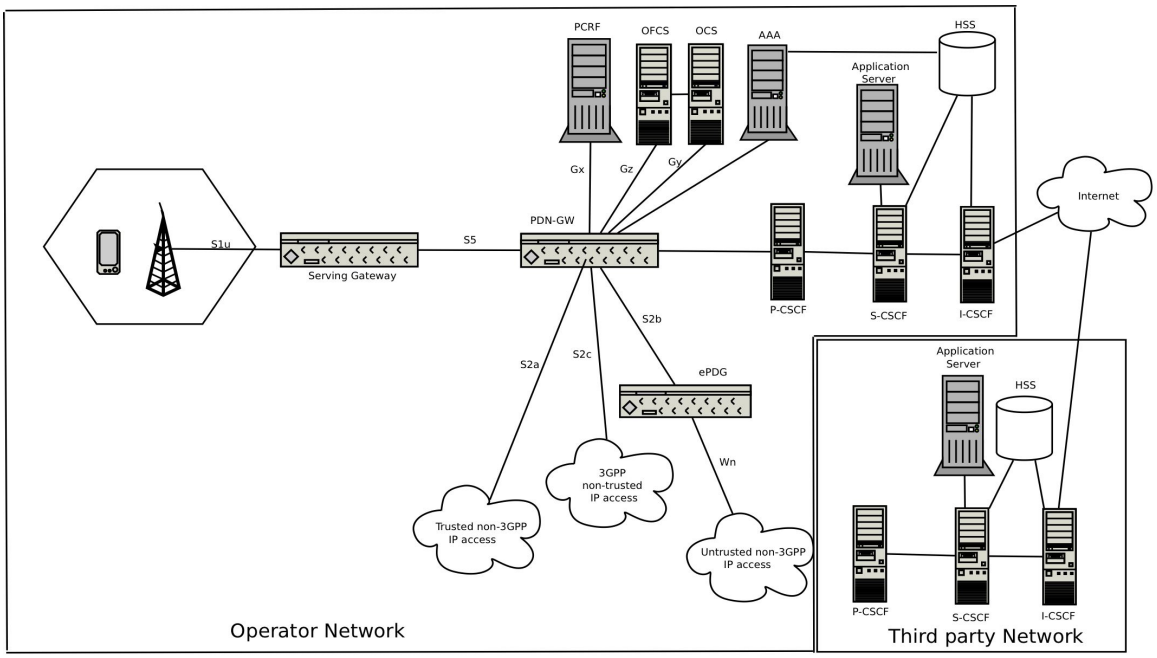



Fig. 1. IMS based integration of operators with trusted third party

```
MESSAGE sip:auctioneer1@3rdparty.com SIP/2.0
Max-Forwards: 70
From: Alice <sip:alice@defaultNetwork.com>; tag=1928301774
To:  3rd Party            <sip:auctioneer1@3rdparty.com>
Call-ID: 1928301774@defaultNetwork.com
CSeq: 1123122 MESSAGE

Content-Disposition: session bidding request
Content-Type: text/plain
Content-Length: xxx
<?xml version="1.0" encoding="UTF-8" ?>
<TypeOfService>VoIP Call</TypeOfService>
<ServiceQualityPreference>Fair</ServiceQualityPreference>
```

Fig. 2. SIP MESSAGE method for user identifying her request, preference over quality, and identity information

indicates the status of registration process. Basically the ACK methods in this case may be an OK or anyother error message. It keeps track of the registration and their acknowledgements using Command Sequence (CSeq) header file.

*UE functional entity:* Given the user has successfully performed SIP registration with IMS platform of the default network. Now if UE wants to conduct a session as per proposed mechanism then she must include the type of service, her preferences and identity in regulation of allowed XML syntax and send this information in the body of SIP MESSAGE to the third party. Here we assume that SIP URI (Uniform Resource Identifier) of the third party is known to the user as part of the contract. A user sends only one session request in one SIP message. UE is also enabled to parse the information received as the part of response that is sent by the third party against her request. The response can basically consequence in accepted or blocked service.

*2) OPNET simulation setup:* In this section, we describe the simulation setup for the proposed network selection approach. In order to simulate the reference scenario presented in Figure 3, the following entities are implemented: i) impairment entity - we developed an impairment entity that introduces specified packet delays, packet losses and is also able to limit the bandwidth shaping using token bucket algorithm. ii) LTE radio access network (eNodeB)), iii) User Equipment(UE), iv) Serving Gateway (S-GW), v) Packet Data Network Gateway (PDN-GW, whereas the following entities used in the simulation are OPNET standard node models: i) Wireless LAN access point, ii) Application server, iii)Ethernet link, iv) Routers, and v) Mobility model.

Note: The packet delay values in simulation only include codec delays(for real-time applications) and transport network delay excluding fixed delay components e.g., equipment related delays, compression and decompression delays etc.

For real-time VoIP applications, we use GSM EFR, G.711, and G.729 codecs in simulation setup, the purpose of using different codecs enable operators to extend offers of different QoE to the users and analyze the users reaction to different offers. For real-time video applications, we use PSNR as video quality metric and make use of EvalVid [27] framework for video quality evaluation. In this setup packet losses are injected using Bernoulli distribution and we use playout buffer of $250ms$ during the reconstruction of video file. We consider a reference video sequence called Highway for this work. The motivation to use this video sequence its repeated reference in a large number studies in video encoding and quality evaluation e.g., Video Quality Experts Group[28]. This video sequence has been encoded in $H.264$ format using the JM codec [29] with CIF resolution $(352 \times 288)$ using a target bit

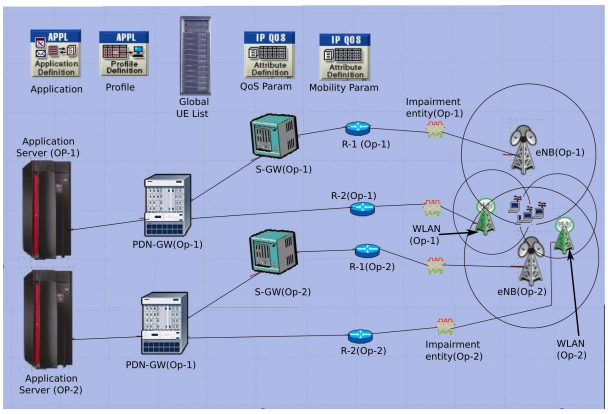



Fig. 3. OPNET Simulation scenario

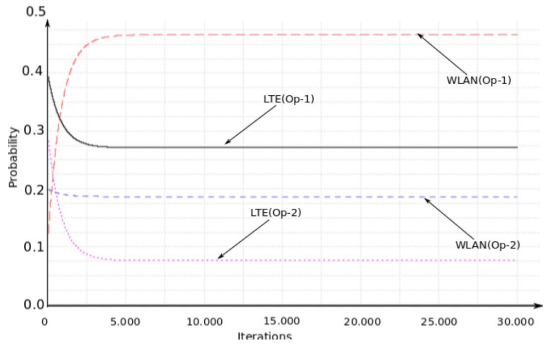



Fig. 4. Evolution of randomized actions for underloaded configuration

rate of $256kbps$. $H.264$ codec has been selected because its widespread use can be seen in future communication devices. The reference video sequence has total $2000$ frames and frame rate of $30fps$. Key frame is inserted after every $10th$ frame which provides good error recovery capabilities. An excellent video quality is indicated by $38.9dB$ as an average PSNR value of encoded video sequence. The video file is transmitted over the IP network considering MTU size of $1024$ bytes.

For non-real-time FTP applications simulation setup, file size is considered to be 20MB, which can be downloaded through LTE and WLAN access network. The choice of file size is dictated by the facts; a) slow start effect of TCP can be ignored, b) correlation of TCP throughput and distribution of packet losses within a TCP can be reduced. Here a bandwidth shaping of $8Mbps$ is performed. We use TCP flavor New Reno with receiver buffer size of $64KB$.

As can be viewed in the Figure-3 that the users under consideration are covered by the two access networks namely LTE and WLAN of two different operators. The integration of these access technologies follow 3GPP recommendations for integration of 3GPP and non-3GPP access technologies [30]. To have greater control of environment in terms of analysis, impairment entities are placed in the transport networks of each access technology. Since the mobility is host-based, therefore MIPv6 based mobility management is implemented at PDN-GW, however for network-based mobility PMIPv6 can be implemented, where LMA resides at ePDG in untrusted integration case. User terminals are multi-interface devices, and are capable of simultaneously connecting to multiple access technologies. We also extensively implement the flow management entity, which acts a relay or apply filter rules over the traffic depending on uplink or downlink traffic.

In order to demonstrate the user-centric based network selection, and demonstrate the effect of learning in such a telecommunication landscape, we run an extensive round of simulation runs. Service requests of different quality classes (user types) are generated by users. The arrival of requests is modeled by Poisson process, and the service class is chosen randomly among voice, data, and video uniformly. The sizes of requests are assumed to be static and are $60kbps$, $150kbps$, and $500kbps$ for voice, data, and video respectively. The capacities of LTE and WLAN network technologies are $32Mbps$ (Downlink)/ $8Mbps$ (Uplink), $8Mbps$ respectively. As the network technologies are owned by two different operators, the technical configuration of the technologies owned by both the operators are very similar. However the service pricing scheme is operator specific, which influences the user-centric network selection decision.

*3) Result Analysis:* Within the simulation settings, we configure that all the users in the system have the same initial probability list i.e., $0.4$,$0.3$,$0.2$,$0.1$ for LTE (Op-1), LTE (Op-2), WLAN (Op-2), and WLAN (Op-1) respectively. We also configure that operator-1 offers lesser service costs when compared with the operator-2, whereas both the operators charge more on LTE than WLAN network technology. The configuration of the technical indices are the same for both the technologies and both the operators, thus the operators offer of technical parameters are influenced by the congestion, available bandwidth, wireless medium characteristics etc. The simulation was run for number of iterations and the convergence of user probabilities of network selection was observed for variable learning schemes. First we analyze the behavior of a fair user in the given settings, as can be observed in Figure 4 that a fair user adjusts its probabilities in the given configuration. As expected the user strategies converge (within relatively small number of iterations) so that she prefers the relatively less costly WLAN (OP-1) more than anyother technology, the probability values of other strategies are the consequences of both technical and non-technical offers of the operators. It should be noted that the Figure 4 is result in underloaded system configurations. i.e., both the technologies of both the operators are under utilized. We now analyze the fair user behavior in the congested system configuration (congested system may defined as the system, where most of the resource are already utilized and the option window of user is squeezed), the results for such configuration are presented in Figure 5. The impact of congestion over the network selection can be seen by strategy convergence of the user. LTE (Op-2) turns out to be the only under loaded network technology, this shrinks the options of the user and hence the different convergence result than that of under-loaded configuration even though the simulation settings remain the similar in both the configurations. These results confirm the superiority of the proposed learning approach in user-centric 4G heterogeneous wireless network selection paradigm. A number of simulations were run and various other results

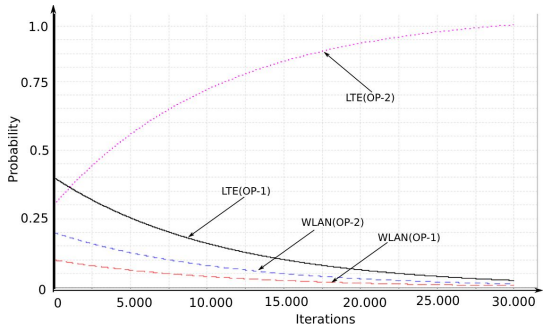



Fig. 5. Evolution of randomized actions for congested configuration

in the similar fashion were taken, where service costs were varied, medium impairments (customized impairments were introduced in the wireless medium with the help of impairment entity) were introduced in the wireless access networks of different operators. The objective of these scenarios was to analyze the behavior of user decision under various dynamics of the system. All the results follow the similar behavior as the ones shown in Figures 4,5 in different configurations. Thus on the basis of the presented results we can confidently claim that the proposed learning scheme fits well to the future user-centric wireless networks paradigm.

### B. Frequency selection and access control

In this subsection we give illustrative example of random medium access control in wireless networks. In wireless communication networks, Medium Access Control (MAC) schemes are used to manage the access of active nodes to a shared channel. As the throughput of the MAC schemes may significantly affect the overall performance of a wireless network, careful design of MAC schemes is necessary to ensure proper operation of a network. Recall the basic rule of slotted Aloha scheme: *if more than two users transmit then there is collision*. Following the idea, one can introduce frequency selection case: *if more than two users transmit at the same time with the same frequency then there is collision*.

We consider $n$ users and $m$ frequencies. $\mathcal{N} := \{1, 2, \ldots, n\}$ is the set of users, $n$ is the total number of users in the system. $\mathcal{F} = \{1, 2, \ldots, m\}$ the set of frequencies for the $n$ users. Each user can choose only one among the $m$ frequencies. Denote by $x_{j,t}(f)$ the probability that user $j$ chooses the frequency $f$ at time $t$. The success probability of user $j$ is given by

$$u_j(x_t) = \sum_{f=1}^{m} x_{j,t}(f) \prod_{j' \neq j} (1 - x_{j',t}(f)).$$

This says that a user $j$ with frequency $f$ has successful transmission only if no other user is using the same frequency. We examine two cases: (i) $m < n$ (ii) $m \geq n$. The state $w$ corresponds to ON/OFF. The state ON means the interface is working and the state OFF means the interface is not working. When the interface is OFF the user cannot access, therefore we look at the probability for the interface to be ON and multiply the performance index by this probability. In the analysis we omit this probability.

*Global optimization:* The global optimization problem consists to maximize the probability of successful transmission of all the system. The problem can be formulated as follows:

$$\begin{cases} \max_x & \sum_{j \in \mathcal{N}} u_j(x) \\ & \forall\ j \in \mathcal{N},\ \sum_{f \in \mathcal{F}} x_j(f) = 1 \\ & \forall\ j \in \mathcal{N},\ \forall f \in \mathcal{F},\ x_j(f) \geq 0 \end{cases}$$

We denote by $\Delta(\mathcal{F}) = \{z,\ \sum_{f \in \mathcal{F}} z_j(f) = 1,\ \forall f,\ z_j(f) \geq 0\}$ the simplex. Then, $\forall j,\ x_j \in \Delta(\mathcal{F})$.

- If $n \leq m$, a direct affectation solves the problem. This implies that we have an exponential number of solutions.
- If $n > m$, affect $m - 1$ of the frequencies to $m - 1$ users. The remaining $n - m + 1$ users remains with one frequency. We have again an exponential number of solutions.

*Equilibrium analysis:* Define a one-shot game given by the collection $\mathcal{G} = (\mathcal{N}, (u_j(.)_{j \in \mathcal{N}}, \mathcal{F})$. We say that $x$ is an equilibrium of $\mathcal{G}$, if

$$\forall j,\ u_j(x) \geq u_j(x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n),\ \forall y_j \in \Delta(\mathcal{F}).$$

We first remark that the above solutions of the optimization problem are pure equilibria of the one-shot game $\mathcal{G} = (\mathcal{N}, (u_j(.))_{j \in \mathcal{N}}, \mathcal{F})$. In particular the global optimum value can be obtained as an equilibrium payoff i.e the so-called *Price of Stability* is one.

There are many other equilibria of the game $\mathcal{G}$. To see this, consider the case where $n > m$. Any configuration where all the frequencies are used and any other strategies of the remaining users is an equilibrium of $\mathcal{G}$.

*Fairness:* When $n > m$ the global optimum and the pure equilibrium payoffs are not fair in the sense that some of the users get 1 and some other 0. A more fair solutions can be obtained using mixed strategies. For example if $\forall\ j, \forall\ f,\ x_j^*(f) = \frac{1}{m}$, the expected payoff of each user is $(1 - \frac{1}{m})^{n-1} > 0$ and the total system payoff is $n(1 - \frac{1}{m})^{n-1}$.

Pareto optimality is a measure of efficiency. An outcome of the game $\mathcal{G}$ is Pareto optimal if there is no other outcome that makes every user at least as well off and at least one user strictly better off. That is, a Pareto Optimal outcome cannot be improved upon without hurting at least one user.

**Lemma 2.** *The above strategy profile $x^*$ is Pareto optimal.*

The proof of this Lemma follows from the fact the strategy maximizes the weighted sum of payoff of the users.

*Learning efficient outcome:* As an illustration we have implemented the Bush-Mosteller based CODIPAS-RL. In Figure-6 we represent the evolution of strategies in a scenario with two users and same action set $m = 2$, $\mathcal{A}_j = \{1, 2\}$ for the two users. As we can observe, the trajectory goes to an equilibrium $(1/2, 1/2)$ which is not efficient. In Figure-7, we represent a convergence to an efficient outcome: global optimum using Bush-Mosteller based CODIPAS-RL for different action sets. Note that, in this scenario the convergence time to be arbitrary close is around 250 iterations which is relatively fast.

*1) Algorithm:* The algorithm CODIPAS-RL is described as follows.

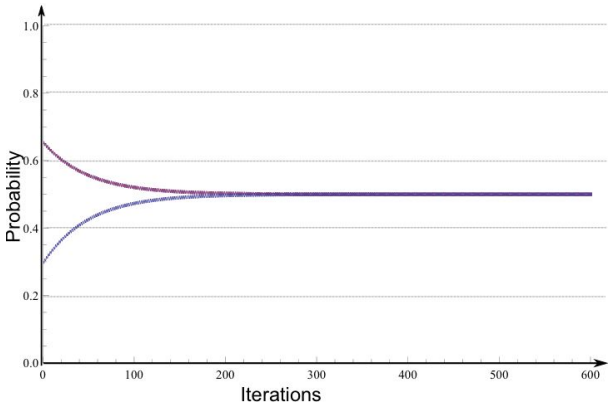



Fig. 6. Convergence to equilibrium

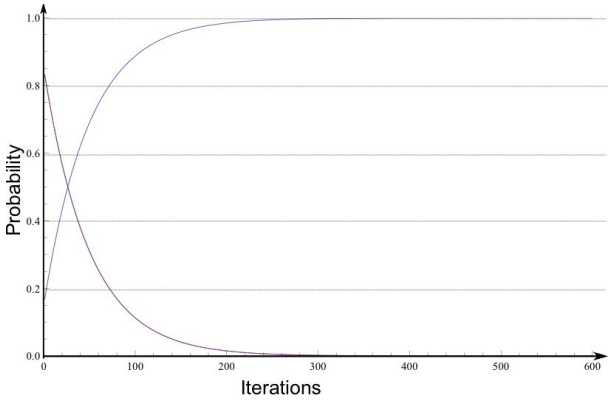



Fig. 7. Convergence to global optimum

**Algorithm 1**: Generic representation of the hybrid CODIPAS-RL

**foreach** *Player* $j$ **do**
  Initial action $a_{j,0}$;
  Initialize to some estimations $\hat{\mathbf{u}}_{j,0}$,;
**end**
**for** *t=1* **to** *max* **do**
  **foreach** *Player* $j$ **do**
    Choose an action $a_{j,t}$ with probability $\mathbf{x}_{j,t}$;
    Observe a numerical value of its noisy payoff $u_{j,t}$;
    Choose one of the learning patterns $l \in \mathcal{L}$ according to $\omega$;
    Update its payoff estimation via $\hat{\mathbf{u}}_{j,t+1}$;
    Update its strategy via $\mathbf{x}_{j,t+1}$;
  **end**
**end**

On the similar lines discussed in the user-centric network selection paradigm, In Figures-8&9, we represent the behavior of the users and their estimated payoff when using variable learning schemes. When the users are active, they can select one of the CORDIPAS learning schemes among $\mathcal{L}_1 - \mathcal{L}_5$ with probability distribution $[1/5, 2/5, 1/5, 1/10, 1/10]$. The users are active with probability 0.9. We choose $\lambda_t = \frac{1}{(t+1)\log(t+1)}$. The Figure-8 represent the evolution of strategies and the Figures-9 represent the estimated payoff evolutions of user

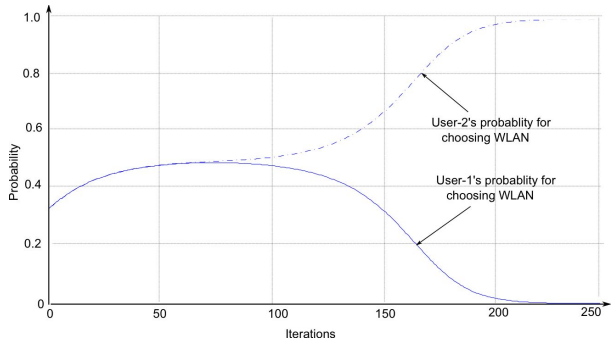



Fig. 8. Evolution of randomized actions

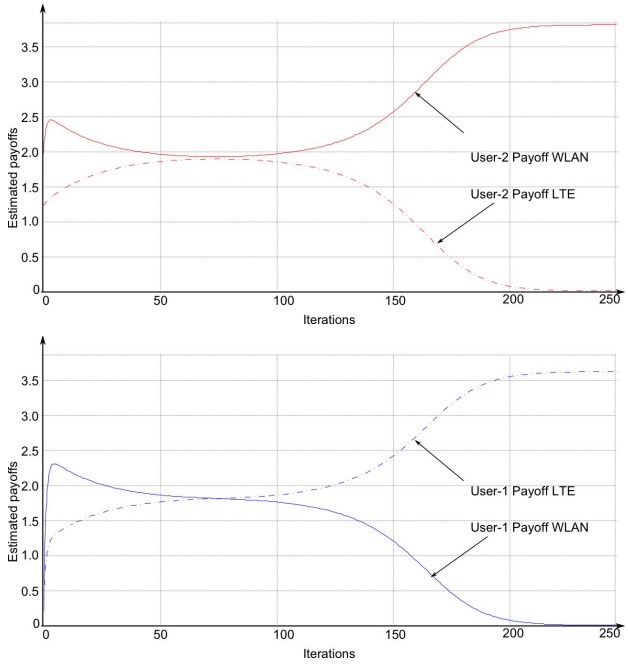



Fig. 9. Evolution of estimated payoffs

1 and 2. As we can observe, the convergence occurs even for random updating time and hybrid CODIPAS-RLs. Not surprisingly, the convergence time seems very large.

At this point it is important to mention that in addition to equilibrium analysis, we have established a convergence to a global optimum for our specific 4G network selection problem. To be best to the authors knowledge, very little is known for the convergence to a *global optimum* in a fully distributed learning way (no coordination, no message exchange, only a numerical noisy and delayed measurement of own payoff is observed). Thus, this is very promising result for extension to other specific classes of wireless games. Moreover, using our analysis, the speed of convergence can be improved by choosing constant learning instead of diminishing learning rates. In that case, a weak convergence can be established. After that, one can conduct the same analysis for the resulting hybrid evolutionary game dynamics. Finally, the dynamic nature of emerging wireless networks allow one to study the importance of time delays, noisy measurement, imperfectness and random number of interaction. The delays can be avoided under appropriated time-scaling. However, for time delay

that are learning rate-dependent, *delayed evolutionary game dynamics may arise as asymptotic pseudo-trajectories* [23].

*Discussions: Fastest learning algorithm*

In this section we address of speed of convergence and running time of simple classes of learning algorithms. The running time analysis is a familiar problem in learning in games as well as in machine learning.

In order to introduce the problem of convergence time, we first start by a classical problem in statistics: Given a target population, how can we obtain a representative sample?

In the context of learning in games, this question can be seen as: Given a list of measurements (such as perceived payoffs), can we obtain a useful information such as best-response strategy or expected payoff distribution?

We consider the class of CODIPAS-RL schemes that generate irreducible aperiodic Markov chain. Let $\mathbf{x}_t$ be an irreducible aperiodic Markov chain with invariant probability distribution $\pi$, having support $\Omega \subseteq \prod_{j\in\mathcal{N}} \mathcal{A}_j$ and let $\mathbb{L}^t$ denote the distribution of $\mathbf{x}_t|\mathbf{x}_0$ for $t \geq 1$. that is

$$\mathbb{L}^t(x,\Gamma) = \mathbb{P}\left(x_t \in \Gamma \mid x_0 = x\right).$$

Then, given any $\epsilon > 0$, can we find an integer $t^*$ such that

$$\parallel \mathbb{L}^t(x,.) - \pi \parallel_{tv} \leq \epsilon,\ \forall t \geq t^*$$

where $tv$ denotes the total variation norm.

Note that, under the above assumptions, $\parallel \mathbb{L}^t(x,.) - \pi \parallel_{tv}$, is non-increasing in $t$. This means that for every draw past $t$ will also be within a range $\epsilon$ from $\pi$, thus providing a representative sample if we keep only the draws after $t^*$. For the Gibbs distributions/Glauber dynamics, there is an enormous amount of research on this problem for a wide variety of Markov chains leading a class of learning schemes in games. Unfortunately, there is apparently little that can be said generally about this problem so that we are forced to analyze each learning scheme chain individually or at most within a limited class of models or situations such as potential, geometric etc.

To simplify the analysis we focus on the reversible Markov chain case, this is for example satisfied by Boltzmann-Gibbs-based CODIPAS-RL. If $\mathbb{L}_{a,a'}(.)$ denotes the transition matrix and $m = \prod_{j\in\mathcal{N}} |\mathcal{A}_j| = |\mathcal{F}|^n$ the number of action profiles, it is well-known that the convergence time to reach the stationary distribution is governed by the second highest eigenvalue [31], [32] of the matrix $(\mathbb{L}_{a,a'})$ after the eigenvalue 1, Let

$$1 = eig_1(\mathbb{L}) \geq eig_2(\mathbb{L}) \geq \ldots \geq eig_m(\mathbb{L}) \geq -1.$$

The speed of convergence is given by the $\frac{1}{1-eig_2(\mathbb{L})}$. The smaller $eig_2(\mathbb{L})$ is the faster the Markov chain $\mathbf{x}_t$ approaches $\pi$.

Based on this observation we define the fastest learning algorithm along the class satisfying the above assumptions as following:

$$\inf_{\mathbb{L}(.)\geq 0} eig_2(\mathbb{L}) \tag{27}$$

$$\pi_a \mathbb{L}_{a,a'} = \pi_{a'} \mathbb{L}_{a',a} \tag{28}$$

$$\sum_{a'\in\mathcal{A}} \mathbb{L}_{a,a'} = 1,\ \forall a \in \Omega. \tag{29}$$

This an optimization problem over the class of learning schemes. Since $eig_2(.)$ is continuous and the set of possible transition matrices constraint is compact, there is at least one optimal transition matrix; the $\inf$ can be by $\min$ i.e an optimal (for the convergence time to $\pi$) learning scheme among the class of CODIPAS satisfying the above assumptions exists.

Since we have the existence result, we need to explain how to find this optimal CODIPAS algorithm. This leads to the question of solvability of (27). Since the eigenvalue $eig_1(.) = 1$ with eigenvector $(1, 1, \ldots, 1)$, we can write the eigenvalue $eig_2(\mathbb{L})$ as an optimization of a quadratic term over vectors:

$$eig_2(\mathbb{L}) = \sup\{\langle v, \mathbb{L}v\rangle \mid \sum_{a\in\Omega} v_a = 0,\ \parallel v \parallel \leq 1\}$$

As a consequence of [31], [32], the convergence time for CODIPAS to be within a range $\epsilon$ to $\pi$ is less than $c(m\log m + m\log(\frac{1}{\epsilon}))$, $c > 0$.

## VI. Concluding remarks

We have presented hybrid and heterogeneous strategic learning schemes in dynamic heterogeneous 4G networks. We have illustrated how important these learning schemes are in wireless systems where the measurement can be imperfect, noisy and delayed and the environment random and changing. Our results are validated through Mathematica numerical examples and OPNET simulations for different service classes over LTE, and WLAN technologies taking into consideration the effect of switching costs in the payoff function. We illustrated the proposed cost of learning CODIPAS-RL scheme to find the corresponding solution in an iterative fashion. Our future work is to extend the heterogeneous cost-to-learn algorithm in the context noisy strategy and randomly varying network topologies.

## Appendix

### Proof of the Propositions

#### A. Proof of Proposition 1

Consider the system of CORDIPAS-RL described in sectionIII. Assume the standard assumptions H2-H3 and assume that proportional learning rates (the ratio is relatively similar and non-vanishing). Then, one can write the CODIPAS-RLs in the form of Robbins-Monro's procedure with weighted coefficient and randomly varying number of players. The Robbins-Monro is $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda_t(f(\mathbf{x}_t) + M_{t+1})$ in $\mathbb{R}^d$ for some $d \geq 1$. To do this, we introduce a reference learning as the maximum for the active users at the current time i.e $\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t}, \mu_{j',t})$. Now the learning rate is a random variable. It is easy to see that this random learning rate satisfies the assumption H3 and it satisfies $\lambda_t \geq 0$, Let $g_{j,t}$ is the limiting of the expected value of $\frac{\lambda_{j,t}}{\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t},\mu_{j',t})} \mathbb{1}_{j\in\mathcal{B}^n(t)}$. The function $\bar{g}_{j,t}$ is the limiting of the expected value of $\frac{\mu_{j,t}}{\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t},\mu_{j',t})} \mathbb{1}_{j\in\mathcal{B}^n(t)}$. The function $f_j^{(l)}$ the expected value of $K_j^{1,(l)}$ when $\max_{j'\in\mathcal{B}^n(t)} \max(\lambda_{j',t}, \mu_{j',t})$ goes to zero. $p_{j,t,l}$ is the probability of the event $\{l_{j,t} = l\}$.

• The function $f$ is clearly Lipschitz since the polymatrix payoff entries are finite for any subsets of players.

• $M_{t+1}$ is a martingale difference sequence with respect to the increasing family of sigma-fields $\mathcal{F}_t = \sigma(\mathbf{x}_{t'}, \hat{\mathbf{u}}_{t'}, M_{t'},\ t' \leq t)$ i.e

$$\mathbb{E}\left(M_{t+1} \mid \mathcal{F}_t\right) = 0$$

• $M_t$ is square integrable and there is a constant $c > 0$, $\mathbb{E}\left(\| \mathbb{M}_{t+1} \|^2 \mid \mathcal{F}_t\right) \leq c(1+ \| \mathbf{x}_t \|^2)$ almost surely, for all $t \geq 0$.

• $\sup_t \| \mathbf{x}_t \| < \infty$ almost surely because remains almost surely in the product of simplices times payoff region by construction. Then, the asymptotic pseudo-trajectory is given by the ordinary differential equation (ODE)

$$\dot{\mathbf{x}}_t = f(\mathbf{x}_t),\ \mathbf{x}_0 \ \text{ fixed.}$$

Thus, we can apply the standard approximations developed in Kushner & Clark 1978, which gives that the asymptotic pseudo-trajectory of the hybrid-delayed-CODIPAS-RL can be written in the following form:

$$\left\{ \begin{array}{c} \frac{d}{dt}\mathbf{x}_{j,t}(s_j) = g_{j,t} \sum_{l \in \mathcal{L}} p_{j,t,l} f^{(l)}_{j,s_j}(\mathbf{x}_{j,t}, \hat{\mathbf{u}}_{j,t}), \\ \frac{d}{dt}\hat{\mathbf{u}}_{j,t}(s_j) = \bar{g}_{j,t}\left(\mathbb{E}_{\mathbf{w},\mathbb{B}^n} U_j^{\mathcal{B}^n}(\mathbf{w}, \mathbf{e}_{j,s_j}, \mathbf{x}_{-j,t} - \hat{\mathbf{u}}_{j,t}(s_j))\right) \\ t \geq 0 \\ \mathbf{x}_{j,0} \in \mathcal{X}_j, \hat{\mathbf{u}}_{j,0} \in \mathbb{R}^{|\mathcal{A}_j|}. \end{array} \right.$$

### B. Proof of Proposition 2

The proof follows similar line as in Proposition 1 but using multiple time-scale stochastic approximations.

### C. Sketch Proof of Proposition 3

Now, we provide a sketch proof of Proposition 3. To prove the Proposition 3, we use tools from hybrid evolutionary game dynamics. We want to rely the outcome of dynamics with the equilibria of the expected robust game. (i) Assume that the homogeneous learning are all NSs. Then the zeros of the heterogeneous dynamics of best response of the homogeneous. Thus, they are best response and the resulting dynamics satisfy (NS).

(ii) If the homogeneous are all (PC) then the heterogeneous is PC by summation of positive terms. Thus, the expected game is a potential game and if the potential function serves as Lyapunov in all these dynamics then global convergence holds for the heterogeneous learning.

(iii) The heterogeneous time-scaling leads to a new class of dynamics obtained by composition of the drift terms. This new class of dynamics may be some convergence properties that the homogeneous learning may not have. This proves that the heterogeneity is crucial for the convergence. The results of (i) and (ii) extends to hybrid CODIPAS-RL by taking the sum over all the learning patterns in the support. (iv) If a hybrid of (PCs) contains at least one (NS) then the "non-Nash rest points" are eliminated. This is because such a point cannot be a rest point of the resulting hybrid dynamics. (v) the result (iv) extends to hybrid evolutionary game dynamics with large number of players (possibly continuum), see [33]. This completes the proof.

## REFERENCES


[1] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *Amer. Econ. Rev.*, 93, 2003.

[2] D. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

[3] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[4] J. R. Marden, G. Arslan, and J. S. Shamma. Joint strategy fictitious play with inertia for potential games. *IEEE Trans. Autom. Control*, 54(2), February 2009.

[5] H. P. Young. Learning by trial and error. *Games and Economic Behavior, Elsevier*, 65:626–643, March 2009.

[6] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[7] H. J. Kushner and D. S. Clark. Stochastic approximation methods for constrained and unconstrained systems. *Springer, New York*, 1978.

[8] A. Benveniste, P. Priouret, and M. Metivier. Adaptive algorithms and stochastic approximations. *Springer Applications Of Mathematics Series, 365 pages*, 1990.

[9] D. S. Leslie and E. J. Collins. Convergent multiple timescales reinforcement learning algorithms in normal form games. *The Annals of Applied Probability*, 13(4):1231–1251, 2003.

[10] Taylor and Jonker. Evoltionarily stable strategies and game dynamics. *Mathematical Bioscience*, 40:145–156, 1978.

[11] J. R. Marden, H. Peyton Young, G. Arslan, and J. S. Shamma. Payoff-based dynamics for multi-player weakly acyclic games. *SIAM Journal on Control and Optimization*, 2009.

[12] Y. Jin, S. Sen, R. Guerin, K. Hosanagar, and Z.-L. Zhang. Dynamics of competition between incumbent and emerging network technologies. *NetEcon*, 2008.

[13] M. Manshaei, J. Freudiger, M. Felegyhazi, P. Marbach, and J. P. Hubaux. On wireless social community networks. *IEEE Infocom*, Apr. 2008.

[14] J. Park S. Ren and M. van der Schaar. User subscription dynamics and revenue maximization in communications markets. *IEEE Infocom, Apr.*, 2011.

[15] H. Tembine, E. Altman, R. ElAzouzi, and Y. Hayel. Bio-inspired delayed evolutionary game dynamics with networking application. *Telecommunication Systems Journal, DOI: 10.1007/s11235-010-9307-1.*, 2010.

[16] A.V. Vasilakos and M. Anastasopoulos. Application of evolutionary game theory to wireless mesh networks. *in "Advances in Evolutionary Computing for System Design, Springer*, 66:249–267, 2007.

[17] Markos P. Anastasopoulos, Dionysia K. Petraki, Rajgopal Kannan, and Athanasios V. Vasilakos. Tcp throughput adaptation in wimax networks using replicator dynamics. *IEEE Trans. Syst. Man Cybern. B, Cybern.*, June 2010.

[18] H. Tembine, E. Altman, R. ElAzouzi, and Y. Hayel. Evolutionary games in wireless networks. *IEEE Trans. Syst. Man Cybern. B, Cybern., Special Issue on Game Theory*, June 2010.

[19] H. P. Young. Strategic learning and its limits. *Oxford University Press*, 2004.

[20] H. Tembine, E. Altman, R. ElAzouzi, and W. H. Sandholm. Evolutionary game dynamics with migration for hybrid power control in wireless communications. *47th SIAM/IEEE CDC*, December 2008.

[21] H. Tembine. Dynamic robust games in mimo systems. *IEEE Trans. Syst. Man Cybern. B, Cybern., 99*, 41:990 – 1002, August 2011.

[22] R. Bush and F. Mosteller. Stochastic models of learning. *Wiley Sons, New York.*, 1955.

[23] H. Tembine. Distributed strategic learning for wireless engineers. *Lecture notes, Supelec, 300 pages*, 2010.

[24] H. Tembine, J. Y. Le Boudec, R. ElAzouzi, and E. Altman. Mean field asymptotic of markov decision evolutionary games. *International IEEE Conference on Game Theory for Networks, Gamenets*, 2009.

[25] Manzoor Ahmed Khan and Umar Toseef. User utility function as quality of experience (qoe). In *Proc. ICN'11*, pages 99–104, 2011.

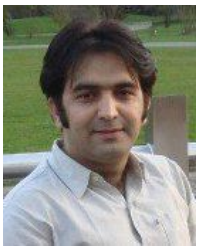
[26] Gonzalo Camarillo and Miguel-Angel Garca-Martn. *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds*. WILEY, 2004.

[27] J. Klaue, B. Rathke, and A. Wolisz. Evalvid - a framework for video transmission and quality evaluation. In *In Proc. 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, pages 255–272, 2003.

[28] Video Quality Experts Group. http://vqeg.org (last accessed september 2, 2010).

[29] S. Shin, S. Bahng, I. Koo, and K. Kim. Qos-oriented packet scheduling schemes for multimedia traffics in ofdma systems. *4th International Conference on Networking*, 2005.

[30] M. La Monaca I. Guardini, E. Demaria. Mobile ipv6 deployment opportunities in next generation 3gpp networks. *16th IST mobile and wireless communications summit Budapest, Hungary*, 2007.
[31] P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of markov chains. *Ann. Appl. Probab.*, 1:36–61, 1991.
[32] Diaconis P. Xiao L. Boyd, S. Fastest mixing markov chain on a graph. *SIAM Rev.*, 46:667–689, 2004.
[33] H. Tembine. Population games in large-scale networks: time delays, mean field dynamics and applications. *LAP, 250 pages*, 2009.



**Manzoor Ahmed Khan** received the Bachelors of Engineering degree in Electronic Engineering from the Mehran University of Engineering and Technology (MUET), Pakistan, in 2001, the MS in computer science degree from Balochistan University of Engineering, Information Technology and Management Sciences, Pakistan, in 2005. He is pursuing his PhD at DAI Labor, Technical University Berlin since 2007. He is the author of several scholarly articles and book chapters. His research interest includes the resource allocation, network selection algorithms in 4G wireless networks, and representation of user Quality of Experience (QoE).

**Hamidou Tembine** is currently assistant professor at Supelec, Gif-sur-Yvette, France. He has been research and teacher assistant at the Computer Science Department, University of Avignon. His main research interests are evolutionary games, population games, mean field stochastic games and their applications. H. Tembine received two Master degrees respectively from Ecole Polytechnique (Palaiseau, France) and University Joseph Fourier, France in 2006. He received the Ph.D degree on *population games with networking applications* from University of Avignon in 2009. He has authored or co-authored over eighty (80) scientific research papers including journals, conferences and workshops.

**Athanasios V. Vasilakos** is currently Professor at the University of Western Macedonia, Greece. He has authored or co-authored over 200 technical papers in major international journals and conferences. He is author/coauthor of five books and 20 book chapters in the areas of communications. Prof. Vasilakos has served as General Chair, Technical Program Committee Chair for many international conferences. He served or is serving as an Editor or/and Guest Editor for many technical journals, such as the IEEE TRANSACTIONS ON NETWORK AND SERVICES MANAGEMENT, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS- PART B: CYBERNETICS, the IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE,ACM TRANSACTIONS ON AUTONOMOUS AND ADAPTIVE SYSTEMS, the IEEE JSAC special issues of May 2009,Jan 2011,March 2011, the IEEE Communications Magazine, ACM/Springer Wireless Networks (WINET), ACM/Springer Mobile Networks and Applications (MONET).He is founding Editor-in-Chief of the International Journal of Adaptive and Autonomous Communications Systems (IJAACS,http://www.inderscience.com/ijaacs) and the International Journal of Arts and Technology (IJART, http://www.inderscience.com/ijart). He is General Chair of the Council of Computing of the European Alliances for Innovation.