Random Multiple Access With Hierarchical Users

Faeze Heydaryan and Jie Luo, *Senior Member, IEEE*

***Abstract*— Classical random access protocols support prioritized user groups by allowing a high priority user to transmit with a relatively high probability. However, when a high priority user competes for the channel with a large number of low priority users, transmission success probability of the high priority user can still diminish to zero. In this paper, a distributed medium access control (MAC) framework is proposed to support hierarchical user groups in a random multiple access system in the following senses. First, when the number of primary users is small, the MAC framework guarantees that channel availability should stay above a pre-determined threshold no matter how many secondary users are competing for the channel. Second, when the number of primary users is large, the MAC framework drives transmission probabilities of the secondary users to zero but does not reject channel access to the primary users. These properties are achieved in a distributed environment without direct message exchange between users, without knowledge on the number of users, and without knowing whether each transmission should belong to a primary or a secondary user. The MAC framework is also extended to systems where each user can be equipped with multiple transmission options.**

***Index Terms*— Random access, hierarchical users, distributed algorithm, wireless network.**

I. Introduction

WITH the rapid growth of mobile computers, digital handheld devices, and smart sensors, traffic in wireless networks is becoming increasingly fragmental, featuring large numbers of bursty short data packets. Due to the difficulty of coordinating a large number of wireless users, a growing proportion of packets are transmitted using distributed medium access control (MAC) protocols where users access a shared wireless channel opportunistically and adjust their transmission schemes individually. Typical examples of such distributed MAC protocols include the 802.11 distributed coordination function (DCF) [1] and its variations, which are widely adopted in WiFi systems as well as in other extended wireless networks. A distributed MAC protocol can often be regarded as the integration of three key components, namely, a random access scheme, a fast adaptation algorithm, and a random scheduling approach. The "random access scheme" regulates how users with short packets should access the shared channel opportunistically. In 802.11 DCF, when a short packet becomes available at a user, the user can transmit the packet opportunistically only if the channel is first sensed to be idle. The conditional packet transmission probability of each user is guarded by an associated backoff window [1]. Depending on the success/failure status of each transmission, which is fed back by the receiver, the corresponding user should follow the "fast adaptation algorithm" to adjust the size of its backoff window accordingly. When a long message becomes available at a user, on the other hand, the "random scheduling approach" is invoked to schedule the transmission. In the example of 802.11 DCF, the user and the receiver should first exchange RTS and CTS handshake packets using the random access scheme. Once the handshake is successful, other users hearing the handshake should remain idle and the channel should then be reserved exclusively for the long message transmission. A detailed introduction and a comprehensive performance analysis of the 802.11 DCF protocol can be found in [2].

To enable theoretical analysis that focuses only on the "random access scheme," a simplified distributed MAC algorithm often assumes equal-sized short packets and assumes that each user should transmit a packet opportunistically according to an associated probability parameter [3]. Under the assumptions of homogeneous users and saturated message queues, if the number of active users is known, optimal transmission probability of the users can be derived theoretically [4], [5]. However, such a result only has limited practical significance because the joint assumptions of saturated message queues and a known number of users is incompatible with the practical environment of bursty packet arrivals. Alternatively, if one regards bursty packet arrivals as users occasionally joining/exiting the system, then assuming saturated message queues at an *unknown* number of active users strikes a reasonable balance between theoretical modeling and practical consideration.[1] Under the additional assumption of incremental transmission probability adaptation, the distributed MAC algorithm falls into a classical stochastic approximation framework where rich mathematical tools are available for its equilibrium and convergence analysis [6]–[8]. For example, in [9], a distributed MAC framework was developed for time-slotted random multiple access over a collision channel. Convergence to the designed system equilibrium is proven based on stochastic approximation analysis. Interestingly, in the case of homogeneous users, the ideal solution of all users having the optimal transmission probability is not included in the set of viable equilibrium

[1]This assumes that tracking the dynamic variation of the number of active users should be a task for the "fast adaptation algorithm."

Manuscript received September 19, 2020; revised January 29, 2021; accepted April 21, 2021. Date of publication May 3, 2021; date of current version October 11, 2021. This work was supported by the National Science Foundation under Grant CNS-1618960. The associate editor coordinating the review of this article and approving it for publication was L. Giupponi. *(Corresponding author: Jie Luo.)*

The authors are with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523 USA (e-mail: faezeh66@colostate.edu; rockey@colostate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3075451.

Digital Object Identifier 10.1109/TWC.2021.3075451

design choices [9]. While the design of 802.11 DCF did not strictly follow the fundamental guidance presented in [9], convergence issues of 802.11 DCF that lead unfortunate users to “deep backoff” states have been widely reported in the literature [10]–[12].

Classical MAC protocols assume that a data link layer user should only determine whether to transmit a packet or not [2], [3], [5], [9]. Such binary transmission/idling decisions do not allow the exploitation of advanced wireless capabilities such as rate, power, and antenna adaptations at the data link layer. To address this architectural inefficiency problem, a new physical layer channel coding theory was developed in [13], [14] to support the distributed communication model by allowing each physical layer transmitter to choose its channel code without sharing the coding decision with other transmitters or with the receiver. The receiver should decode the messages if a pre-determined reliability requirement can be met, and should report collision otherwise. An achievable region was defined in the space of the coding choices of the users in the following senses. If the coding choices happen to locate inside the region, the receiver should detect it and should decode the messages reliably. If the coding choices happen to locate outside the region, the receiver should report collision reliably. Fundamental limit of the multiple access channel was characterized by the maximum achievable region, termed the “distributed capacity” [15]. It was shown that the distributed capacity coincides with the Shannon capacity region of the multiple access channel, in a sense explained in [14], [15]. As also explained in [14], [15], the new channel coding theory enabled a physical-link layer interface enhancement that supports multiple transmission options at each link layer user. It also enabled the derivation of a link layer channel model from the physical layer channel and coding details of the data packets.

Following the development of the distributed channel coding theory, in [15], [16], a distributed MAC framework was proposed for a time-slotted multiple access network with homogeneous users and with saturated message queues. This MAC framework extended the one presented in [9] in the following aspects. First, the MAC framework allows each link layer user to be equipped with multiple transmission options. A transmission probability vector is associated with each user to characterize the respective transmission probabilities using the available options. Second, the MAC framework supports a general link layer channel model that can be derived from the physical layer channel based on the distributed channel coding theorems. Third, the MAC framework supports the maximization of a general symmetric network utility. [15], [16] also proposed to measure contention level of the multiple access channel using the reception probability of a carefully designed virtual packet. This is an extension to the classical approach of measuring channel contention level using the channel idling probability [9]. It was shown that the distributed MAC algorithm can help all users adapting their transmission probability vectors to a pre-designed equilibrium. Similar to the observation reported in [9], due to convergence conditions, usually the designed equilibrium can only be near optimal, as opposed to being exactly optimal, with respect to the chosen utility [15], [16].

Diversity of wireless devices and applications often requires wireless networks to provide differentiated services to users in the sense of supporting user groups with different priority levels. While many approaches have been proposed to support priority users in various networks [17]–[19], most of them share a similar basic idea with a common property described below. Take the enhanced DCF (EDCF) protocol in 802.11e for example [20]. Users (or traffics) in 802.11e EDCF can be assigned to four different priority levels with different adaptation schemes on their backoff windows. A high priority user generally maintains a backoff window smaller in size than that of a low priority user. Consequently, when messages are available, the transmission probability of a high priority user is always larger in value than that of a low priority user. This gives high priority users an advantage over low priority users in getting their packets through the shared wireless channel. However, such a priority user structure is “soft” in the sense that, when the system has a large number of users whose transmission activities cause a significant level of contention, the packet transmission success probability of each user can still be driven down close to zero irrespective of the priority level of the user.

Recent trend of dynamic spectrum access (DSA) created the new demand of supporting a “hard” hierarchical user structure in wireless systems [21], [22]. Take channel sharing with the primary-secondary user structure for example. It is expected that secondary users should access the channel only if they can guarantee no disturbance to communication activities of the primary users. Existing DSA literature often assumes that secondary users should be able to identify whether an existing transmission should belong to a primary user or not. Disruptive interference is often avoided with online coordinations between primary and secondary users or within the secondary user group. There is little discussion on how to support hierarchical user groups in a random access environment, where users do not exchange information with each other directly, and where packet collision is part of the natural transmission outcomes. One may mistakenly think that a “hard” hierarchical user structure can be established by considering the scheduled long message transmissions alone. Unfortunately, because long message transmissions in a distributed MAC protocol need to be scheduled first by exchanging handshake packets using the random access scheme [1], [2], [20], it is therefore necessary to enforce a hierarchical user structure in the random access scheme to guarantee channel access for the primary users. In fact, it is possible to argue that, in a distributed MAC protocol, both “hard” hierarchical user structure and “soft” priority user structure can be established *only* in the random access scheme, i.e., without introducing further user differentiation in the scheduled transmissions. This is because if a user structure can be established by differentiating the transmission success probabilities of short packets from different user groups, then such a structure naturally extends to the long message transmissions through its impact on the success probabilities of the RTS and CTS handshake packets.

In this paper, we extend the distributed MAC framework proposed in [15], [16] to support hierarchical user groups in a time-slotted random multiple access system. The "hard" hierarchical user structure is established in the following senses. First, when the number of primary users is small, the MAC protocol guarantees that contention level of the channel at the equilibrium should stay below a pre-determined threshold no matter how many secondary users are competing for the channel. Second, when the number of primary users is large, the MAC protocol guarantees that transmission probability of each secondary user at the equilibrium should be driven down to zero. However, the MAC protocol does not reject channel access to any primary user even though transmission activities of the primary users naturally lead to a low packet transmission success probability. To the best of our knowledge, a distributed MAC algorithm that achieves such a property is not yet available in the literature. We introduce the distributed MAC framework first for random access systems with each user only having a single transmission option. The MAC framework is then extended to systems where each user is equipped with multiple transmission options. Simulation results are provided to demonstrate performances and properties of the distributed MAC algorithms under various system settings.

To help reading the technical contents of the paper, we summarize the definitions of a list of key variables below.

Definitions of Key Variables

K_p, K_s, K : actual numbers of primary, secondary, and all users.

$\hat{K}$: estimated number of users.

q_v : actual channel contention measure, which is the success probability of the designed virtual packet.

q_v^* : theoretical channel contention measure.

When each user is equipped with a single transmission option

p_k : transmission probability of user k.

$\boldsymbol{p}$: a vector that contains the transmission probabilities of all users.

$\boldsymbol{p}^*$: the vector of transmission probabilities of all users at an equilibrium.

$\{C_{rj}\}$: real channel parameter set. C_{rj} is the conditional success probability of a real packet should it be transmitted in parallel with j other real packets.

$\{C_{vj}\}$: virtual channel parameter set. C_{vj} is the success probability of the virtual packet should it be transmitted in parallel with j real packets.

x^* : the limit of Kp as $K \to \infty$. The value of x^* is obtained from the optimization of the asymptotic utility function.

When each user is equipped with multiple transmission options

$\boldsymbol{p}_k = p_k \boldsymbol{d}_k$: transmission probability vector of user k. p_k is the probability that user k transmits a packet. $\boldsymbol{d}_k$, termed the transmission direction vector of user k, specifies the probabilities of user k choosing the corresponding transmission options given that user k transmits a packet.

$\boldsymbol{P}$: a macro vector that contains the transmission probability vectors of all users.

$\tilde{\boldsymbol{P}}$: target transmission probability vectors of all users computed using *noisy* measurements.

$\hat{\boldsymbol{P}}$: theoretical target transmission probability vectors of all users computed using *noiseless* measurements.

$\boldsymbol{P}^*$: transmission probability vectors of all users at an equilibrium.

$\{C_{rij}(\boldsymbol{d})\}$: real channel parameter function set. Under the assumption that all users have the same transmission direction vector $\boldsymbol{d}$, $C_{rij}(\boldsymbol{d})$ is the conditional success probability of a real packet with the ith transmission option should it be transmitted in parallel with j other real packets.

$\{C_{vj}(\boldsymbol{d})\}$: virtual channel parameter function set. Under the assumption that all users have the same transmission direction vector $\boldsymbol{d}$, $C_{vj}(\boldsymbol{d})$ is the success probability of the virtual packet should it be transmitted in parallel with j real packets.

II. A Stochastic Approximation Framework

Consider a wireless multiple access network with K users (transmitters) and a common receiver. Among the users, K_p of them are labeled as "primary" users and K_s of them are labeled as "secondary" users. $K = K_p + K_s$. Each user only knows its own label, but does not know the labels of other users, as well as the values of K_p, K_s and K. As we will explain later, users will adopt the same type of MAC algorithms, but with users of different labels setting their key parameters differently. Other than such a difference, we assume that the users are homogeneous.

Let time be slotted with each slot equaling the length of one packet. Let M be a positive integer. Assume that each user is equipped with M transmission options plus an idling option, and is backlogged with a saturated message queue. At the beginning of each time slot t, a user should either be idle or randomly choose a transmission option to send a message, with corresponding probabilities being specified by an associated probability vector. Transmission decisions of the users are made individually, and they are shared neither among the users nor with the receiver. The M-length probability vector associated to user k, $k = 1, \ldots, K$, is denoted by $\boldsymbol{p}_k(t)$ for time slot t. We write $\boldsymbol{p}_k(t) = p_k(t)\boldsymbol{d}_k(t)$, with $0 \leq p_k(t) \leq 1$ being the probability that user k transmits a packet in time slot t, and with vector $\boldsymbol{d}_k(t)$ specifying the conditional probabilities for user k to choose each of the transmission options should it decide to transmit a packet [15]. We term $p_k(t)$ the "transmission probability" of user k, and term $\boldsymbol{d}_k(t)$ the "transmission direction" vector of user k.

At the end of each time slot t, based upon available channel feedback, each user k derives a target probability vector $\tilde{\boldsymbol{p}}_k(t)$. User k then updates its transmission probability vector by

$$\boldsymbol{p}_k(t+1) = \boldsymbol{p}_k(t) + \alpha(t)(\tilde{\boldsymbol{p}}_k(t) - \boldsymbol{p}_k(t)), \tag{1}$$

where $\alpha(t) > 0$ is a step size parameter of time slot t. Let $\boldsymbol{P}(t) = [\boldsymbol{p}_1^T(t), \boldsymbol{p}_2^T(t), \ldots, \boldsymbol{p}_K^T(t)\ ]^T$ denote an MK-length vector that consists of the transmission probability vectors of all users in time slot t. Let $\tilde{\boldsymbol{P}}(t) = [\tilde{\boldsymbol{p}}_1^T(t), \tilde{\boldsymbol{p}}_2^T(t), \ldots, \tilde{\boldsymbol{p}}_K^T(t)\ ]^T$ denote the corresponding target vector. According to (1), $\boldsymbol{P}(t)$ is updated by

$$\boldsymbol{P}(t+1) = \boldsymbol{P}(t) + \alpha(t)(\tilde{\boldsymbol{P}}(t) - \boldsymbol{P}(t)). \tag{2}$$

Probability adaptation given in (2) falls into the stochastic approximation framework [6]–[8], where the target probability vector $\tilde{\boldsymbol{P}}(t)$ is often calculated from noisy estimates of certain system variables, e.g., channel idling probability.

Define $\hat{\boldsymbol{P}}(t) = [\hat{\boldsymbol{p}}_1^T(t), \hat{\boldsymbol{p}}_2^T(t), \ldots, \hat{\boldsymbol{p}}_K^T(t)\ ]^T$ as the "theoretical value" of $\tilde{\boldsymbol{P}}(t)$ under the assumption that there is no measurement noise and no feedback error in time slot t. Let $E_t[\tilde{\boldsymbol{P}}(t)]$ be the conditional expectation of $\tilde{\boldsymbol{P}}(t)$ given system state at the beginning of time slot t. The difference between $E_t[\tilde{\boldsymbol{P}}(t)]$ and $\hat{\boldsymbol{P}}(t)$ is defined as the bias in the target probability vector calculation, denoted by $\boldsymbol{G}(t)$.

$$\boldsymbol{G}(t) = E_t[\tilde{\boldsymbol{P}}(t)] - \hat{\boldsymbol{P}}(t). \tag{3}$$

We assume that, given the communication channel, both $\hat{\boldsymbol{P}}(t) = \hat{\boldsymbol{P}}(\boldsymbol{P}(t))$ and $\boldsymbol{G}(t) = \boldsymbol{G}(\boldsymbol{P}(t))$ should only be functions of $\boldsymbol{P}(t)$, which is the transmission probability vector in time slot t.

The following two conditions are typically required for the convergence of a stochastic approximation algorithm [15], [16].

Condition 1 (Mean and Bias): There exists a constant $K_m > 0$ and a bounding sequence $0 \leq \beta(t) \leq 1$, such that $\|\boldsymbol{G}(\boldsymbol{P}(t))\| \leq K_m \beta(t)$, where $\|.\|$ denotes the second order norm. We assume that $\beta(t)$ is controllable in the sense that one can design protocols to ensure $\beta(t) \leq \epsilon$ for any chosen $\epsilon > 0$ and for large enough t.

Condition 2 (Lipschitz Continuity): There exists a constant $K_l > 0$, such that $\|\hat{\boldsymbol{P}}(\boldsymbol{P}_a) - \hat{\boldsymbol{P}}(\boldsymbol{P}_b)\| \leq K_l \|\boldsymbol{P}_a - \boldsymbol{P}_b\|$, for all $\boldsymbol{P}_a, \boldsymbol{P}_b$.

According to stochastic approximation theory [6]–[8], if the above two conditions are satisfied, and values of the step size sequence $\alpha(t)$ and the bounding sequence $\beta(t)$ are small enough, then trajectory of the transmission probability vector $\boldsymbol{P}(t)$ under distributed adaptation algorithm given in (2) can be approximated by the following associated ordinary differential equation (ODE) in a sense explained in [6], [8],

$$\frac{d\boldsymbol{P}(t)}{dt} = -[\boldsymbol{P}(t) - \hat{\boldsymbol{P}}(t)], \tag{4}$$

where we used t again to denote the continuous time variable. Because all entries of $\boldsymbol{P}(t)$ and $\hat{\boldsymbol{P}}(t)$ stay in the range of $[0, 1]$, any equilibrium $\boldsymbol{P}^* = [\boldsymbol{p}_1^{*T}, \ldots, \boldsymbol{p}_K^{*T}]^T$ of the associated ODE must satisfy

$$\boldsymbol{P}^* = \hat{\boldsymbol{P}}(\boldsymbol{P}^*). \tag{5}$$

Suppose that the associated ODE given in (4) has a unique solution at $\boldsymbol{P}^*$, then the following convergence results can be obtained from the standard conclusions in the stochastic approximation literature.

Theorem 1: [15, Theorem 4.1] For distributed transmission probability adaptation given in (2), assume that the associated ODE given in (4) has a unique stable equilibrium at $\boldsymbol{P}^*$. Suppose that $\alpha(t)$ and $\beta(t)$ satisfy the following conditions

$$\sum_{t=0}^{\infty} \alpha(t) = \infty, \sum_{t=0}^{\infty} \alpha(t)^2 < \infty, \sum_{t=0}^{\infty} \alpha(t)\beta(t) < \infty. \tag{6}$$

Under Conditions 1 and 2, $\boldsymbol{P}(t)$ converges to $\boldsymbol{P}^*$ with probability one.

Theorem 2: [15, Theorem 4.2] For distributed transmission probability adaptation given in (2), assume that the associated ODE given in (4) has a unique stable equilibrium at $\boldsymbol{P}^*$. Let Conditions 1 and 2 hold true. For any $\epsilon > 0$, there exists a constant $K_w > 0$, such that, if $\alpha(t)$ and $\beta(t)$ satisfy the following constraint with $0 < \underline{\alpha} < \overline{\alpha} < 1$

$$\exists T_0 \geq 0, \underline{\alpha} \leq \alpha(t) \leq \overline{\alpha}, \beta(t) \leq \sqrt{\overline{\alpha}}, \forall t \geq T_0, \tag{7}$$

then $\boldsymbol{P}(t)$ should converge weakly to $\boldsymbol{P}^*$ in the following sense

$$\limsup_{t \to \infty} Pr\left\{\|\boldsymbol{P}(t) - \boldsymbol{P}^*\| \geq \epsilon\right\} < K_w \overline{\alpha}. \tag{8}$$

Theorems 1 and 2 provided convergence guarantee for a class of distributed MAC algorithms. Within the presented stochastic approximation framework, the key question is how to design a distributed MAC algorithm to satisfy Conditions 1 and 2 and to place the equilibrium of the associated ODE at a desired point.

In each time slot, the receiver makes message recovery and collision detection decisions for the data packets without knowing the transmission status of the users. The receiver also assumes the existence of a virtual packet, and makes a decision on whether virtual packet reception should be regarded as successful or not. Only one virtual packet is assumed in each time slot, and virtual packets assumed in different time slots are identical. As explained in [15], [16], a virtual packet is an assumed packet that is not physically transmitted by any user. Virtual packet detection essentially checks whether current operation point of the channel is inside its fundamental limit with a pre-determined margin. Such detection is supported by the distributed channel coding theory and has been extensively investigated in [13]–[15]. We assume that the receiver should estimate the success probability of the virtual packet, denoted by $q_v(t)$ for time slot t, and feed it back to the users. $q_v(t)$ should then be used by each user to derive its target probability vector. As explained in [15, Section 4], virtual packet is an effective tool to model and to measure the contention level of a link-layer channel. Note that virtual packet design and detection can be simple to implement. For example, if the link-layer multiple access channel is a collision channel, and if we design the virtual packet to have the same coding details of a real packet, then reception of the virtual packet should be regarded as successful if and only if no real packet is transmitted. Virtual packet reception in this case is equivalent to channel idling status detection, which can be done easily. General considerations on virtual packet design are explained in [15, Section 4.2] and will not be discussed in this paper.

As introduced in [15, Section 4.3], we model the link layer multiple access channel using two sets of parameter functions, both can be theoretically derived from the physical layer channel model and coding details of the packets. These parameter functions are defined under the assumption that all users should have the same transmission direction vector $\boldsymbol{d}$. The first function set, $\{C_{rij}(\boldsymbol{d})\}$ for $1 \leq i \leq M$ and $j \geq 0$, is termed the "real channel parameter function set".

$C_{rij}(\boldsymbol{d})$ denotes the conditional success probability of a real packet corresponding to the ith transmission option, should the packet be transmitted in parallel with j other real packets. The second function set, $\{C_{vj}(\boldsymbol{d})\}$ for $j \geq 0$, is termed the "virtual channel parameter function set". $C_{vj}(\boldsymbol{d})$ denotes the success probability of the virtual packet should it be transmitted in parallel with j real packets. We assume that $C_{vj}(\boldsymbol{d}) \geq C_{v(j+1)}(\boldsymbol{d})$ for all $j \geq 0$ and for all $\boldsymbol{d}$, meaning that the virtual packet should not get a better chance to go through the channel if the number of parallel real packet transmissions increases. Both $\{C_{rij}(\boldsymbol{d})\}$ and $\{C_{vj}(\boldsymbol{d})\}$ are assumed to be known at the users as well as at the receiver.

III. Supporting Hierarchical Users With a Single Transmission Option

In this section, we consider the simple scenario when each user is only equipped with a single transmission option. In this case, the associated transmission probability vector of a user, say $\boldsymbol{p}_k$ for user k, is degraded to a scalar variable, denoted by p_k. Parameters used to model the multiple access channel are also degraded to the "real channel parameter set" $\{C_{rj}\}$ and the "virtual channel parameter set" $\{C_{vj}\}$, whose definitions are already introduced above.

Let the system have K_p primary users and K_s secondary users, where the values of K_p and K_s are unknown to the users. Let $K = K_p + K_s$. We term q_v the "channel contention measure," which is the success probability of the virtual packet. If all users have the same transmission probability p, q_v can be calculated by

$$q_v(p, K) = \sum_{j=0}^{K} \binom{K}{j} p^j (1-p)^{K-j} C_{vj}. \tag{9}$$

Without global information, a primary (secondary) user assumes that the system should only contain an unknown number of primary (secondary) users. Upon receiving the feedback of q_v, each primary (secondary) user should obtain an estimated number of users, denoted by $\hat{K}$,[2] and then use $\hat{K}$ to determine the corresponding transmission probability target. Such an operation requires each primary (secondary) user to design two key functions, denoted by $p_p^*(\hat{K})$ and $q_{vp}^*(\hat{K})$ ($p_s^*(\hat{K})$ and $q_{vs}^*(\hat{K})$), both are functions of $\hat{K}$. $p_p^*(\hat{K})$ ($p_s^*(\hat{K})$) represents the designed transmission probability of the user if the multiple access system only contains $\hat{K}$ primary (secondary) users. $q_{vp}^*(\hat{K})$ ($q_{vs}^*(\hat{K})$), on the other hand, represents the "theoretical channel contention measure" if the multiple access system only contains $\hat{K}$ primary (secondary) users, and all users have the same transmission probability of $p_p^*(\hat{K})$ ($p_s^*(\hat{K})$). Both functions should be designed for all values of $\hat{K} \geq 0$, including both integer and non-integer values. Define $K_{p\min}$ ($K_{s\min}$) as the maximum $\hat{K}$ that maximizes $p_{vp}^*(\hat{K})$ ($p_{vs}^*(\hat{K})$).

Given the two key functions, the distributed MAC algorithm should operate as follows. Note that the MAC algorithm is similar to the one presented in [15, Section 4.3], except that primary (secondary) users should use their own designed functions of $p_p^*(\hat{K})$ and $q_{vp}^*(\hat{K})$ ($p_s^*(\hat{K})$ and $q_{vs}^*(\hat{K})$).

Distributed MAC Algorithm:

1) Each user initializes its transmission probability parameter.
2) Let $Q > 0$ be a pre-determined integer. Over an interval of Q time slots, the receiver measures the success probability of the virtual packet, denoted by q_v, and feeds q_v back to all users.
3) Upon receiving q_v, each primary (secondary) user should derive an estimate of the number of users $\hat{K}$ by solving the following equation.

$$q_{vp}^*(\hat{K}) = q_v \qquad (q_{vs}^*(\hat{K}) = q_v). \tag{10}$$

If a $\hat{K}$ satisfying (10) cannot be found, a primary (secondary) user should set $\hat{K} = K_{p\min}$ ($\hat{K} = K_{s\min}$) if $q_v > \max_{\hat{K}} q_{vp}^*(\hat{K})$ ($q_v > \max_{\hat{K}} q_{vs}^*(\hat{K})$), or should set $\hat{K} = \infty$ otherwise.
4) Each primary (secondary) user, say user k_p (k_s), updates its transmission probability by

$$p_{k_p} = (1-\alpha)p_{k_p} + \alpha p_p^*(\hat{K}),$$
$$(p_{k_s} = (1-\alpha)p_{k_s} + \alpha p_s^*(\hat{K}),) \tag{11}$$

where α is the step size parameter for user k_p (k_s).
5) The process is repeated from Step 2 till transmission probabilities of all users converge.

We will now introduce the specific design of the two key functions. Let us first focus on a primary user. We assume that primary users intend to maximize a symmetric network utility, denoted by $U_p(\hat{K}, p_p, \{C_{rj}\})$, under the assumption that the system contains $\hat{K}$ homogeneous primary users and all users have the same transmission probability p_p. Let x_p^* be obtained from the following asymptotic utility optimization.

$$x_p^* = \arg\max_x \lim_{\hat{K}\to\infty} U_p\left(\hat{K}, \frac{x}{\hat{K}}, \{C_{rj}\}\right). \tag{12}$$

Based on the utility optimization objective, a primary user should design its desired transmission probability function $p_p^*(\hat{K})$ as

$$p_p^*(\hat{K}) = \frac{x_p^*}{\max\{\hat{K}, \hat{K}_{p\min}\} + b_p}, \tag{13}$$

where $\hat{K}_{p\min}$ and b_p are design parameters whose values should be determined by following the guideline given in [15, Section 4.2]. Particularly, $\hat{K}_{p\min}$ should take an integer value slightly less than x_p^*, and b_p should be chosen to satisfy $b_p > \max\{1, x_p^* - \gamma_{p\epsilon_v}\}$, where $\gamma_{p\epsilon_v}$ is a parameter defined in [15, Theorem 4.4]. We skip the detailed definition of $\gamma_{p\epsilon_v}$ here because usually a good design should yield $x_p^* - \gamma_{p\epsilon_v} < 1$, and hence the effective constraint on b_p should be simplified to $b_p > 1$ [15, Section 4.2].

With the $p_p^*(\hat{K})$ function given by (13), for integer-valued $\hat{K}$, $q_{vp}^*(\hat{K})$ should equal the actual channel contention measure when the estimated number of users is accurate and when

[2] $\hat{K}$ represents an estimated number of users under the assumption that the system only contains primary (secondary) users. When such an assumption is not correct, the value of $\hat{K}$ may not have a clear practical meaning.

all users have the same transmission probability of $p_p^*(\hat{K})$. In other words, we should have

$$q_{vp}^*(\hat{K}) = q_v(p_p^*(\hat{K}), \hat{K}), \tag{14}$$

where $q_v(p_p^*(\hat{K}), \hat{K})$ can be further calculated using (9). For non-integer-valued $\hat{K}$, the "theoretical channel contention measure" function $q_{vp}^*(\hat{K})$ should be designed using the following linear interpolation approach [15, Section 4.2].

$$q_{vp}^*(\hat{K}) = \frac{p_p^*(\hat{K}) - p_p^*(\lfloor \hat{K} \rfloor + 1)}{p_p^*(\lfloor \hat{K} \rfloor) - p_p^*(\lfloor \hat{K} \rfloor + 1)} q_v(p_p^*(\hat{K}), \lfloor \hat{K} \rfloor) + \frac{p_p^*(\lfloor \hat{K} \rfloor) - p_p^*(\hat{K})}{p_p^*(\lfloor \hat{K} \rfloor) - p_p^*(\lfloor \hat{K} \rfloor + 1)} q_v(p_p^*(\hat{K}), \lfloor \hat{K} \rfloor + 1), \tag{15}$$

where $\lfloor \hat{K} \rfloor$ represents the largest integer below $\hat{K}$.

When $p_p^*(\hat{K})$ and $q_{vp}^*(\hat{K})$ functions are designed according to (13), (14), and (15), we have the following two monotonicity properties. On one hand, given K, the channel contention measure function $q_v(p_p^*(\hat{K}), K)$ is monotonically non-decreasing in $\hat{K}$ [15, Theorem 4.3]. On the other hand, the theoretical channel contention measure function $q_{vp}^*(\hat{K})$ is monotonically non-increasing in $\hat{K}$ and is strictly decreasing in $\hat{K}$ for $\hat{K} \geq \hat{K}_{p\min}$ [15, Theorem 4.4]. The basic considerations behind the design of $p_p^*(\hat{K})$ and $q_{vp}^*(\hat{K})$ functions can be briefly explained as follows [15, Section 4.2]. Under the assumption that the system only contains primary users, setting $p_p^*(\hat{K})$ at $p_p^*(\hat{K}) \approx \frac{x_p^*}{\hat{K}}$ is asymptotically optimal (or close to optimal for large $\hat{K}$) in terms of symmetric utility maximization. It is also a general observation that setting $p_p^*(\hat{K})$ at $p_p^*(\hat{K}) \approx \frac{x_p^*}{\hat{K}}$ should be not far from optimal for all $\hat{K}$ values and for most of the utility functions of interest. If we term $p_p^*(\hat{K}) \approx \frac{x_p^*}{\hat{K}}$ the ideal solution, then the proposed design given in (13) should be close to ideal, with the necessary revisions to achieve the desired monotonicity properties required for the convergence proof of the distributed MAC algorithm [15, Section 4.2].

Next, let us switch focus to a secondary user. A secondary user should also design two key functions, denoted respectively by $p_s^*(\hat{K})$ and $q_{vs}^*(\hat{K})$, both are functions of $\hat{K}$ which represents the estimated number of users under the assumption that the system should only contain secondary users. Differs from the design of a primary user, $p_s^*(\hat{K})$ and $q_{vs}^*(\hat{K})$ functions need to be designed to enforce the hierarchical user structure. We will show later that, a hierarchical user structure can be achieved by raising the tail of the $q_{vs}^*(\hat{K})$ function above a pre-determined threshold $\underline{q_v}$, i.e. by imposing the following constraint.

$$\lim_{\hat{K} \to \infty} q_{vs}^*(\hat{K}) \geq \underline{q_v}. \tag{16}$$

The value of $\underline{q_v}$ should be chosen according to the quality of service requirement of the primary users. To satisfy the constraint, let x_s^* be obtained by solving the following equation.

$$\lim_{\hat{K} \to \infty} q_v\left(\frac{x_s^*}{\hat{K}}, \hat{K}\right) = \underline{q_v}, \tag{17}$$

where $q_v\left(\frac{x_s^*}{\hat{K}}, \hat{K}\right)$ is further defined in (9). A secondary user should design its desired transmission probability function $p_s^*(\hat{K})$ as

$$p_s^*(\hat{K}) = \frac{x_s^*}{\max\{\hat{K}, \hat{K}_{s\min}\} + b_s}, \tag{18}$$

where $\hat{K}_{s\min}$ and b_s are design parameters whose values should be determined by following the guideline given in [15, Section 4.2]. Particularly, $\hat{K}_{s\min}$ should take an integer value slightly less than x_s^*, and b_s should satisfy an effective constraint of $b_s > 1$ [15, Section 4.2]. Similar to the design of a primary user, a secondary user should design the "theoretical channel contention measure" function $q_{vs}^*(\hat{K})$ as

$$q_{vs}^*(\hat{K}) = \frac{p_s^*(\hat{K}) - p_s^*(\lfloor \hat{K} \rfloor + 1)}{p_s^*(\lfloor \hat{K} \rfloor) - p_s^*(\lfloor \hat{K} \rfloor + 1)} q_v(p_s^*(\hat{K}), \lfloor \hat{K} \rfloor) + \frac{p_s^*(\lfloor \hat{K} \rfloor) - p_s^*(\hat{K})}{p_s^*(\lfloor \hat{K} \rfloor) - p_s^*(\lfloor \hat{K} \rfloor + 1)} q_v(p_s^*(\hat{K}), \lfloor \hat{K} \rfloor + 1), \tag{19}$$

where $\hat{K}$ can take both integer and non-integer values.

According to [15, Theorem 4.3], given K, $q_v(p_s^*(\hat{K}), K)$ should be monotonically non-decreasing in $\hat{K}$. Meanwhile, according to [15, Theorem 4.4], $q_{vs}^*(\hat{K})$ should be monotonically non-increasing in $\hat{K}$ and should be strictly decreasing in $\hat{K}$ for $\hat{K} \geq \hat{K}_{s\min}$. Due to (17), we have $q_{vs}^*(\hat{K}) \geq \underline{q_v}$ for all $\hat{K}$.

With the above design, the distributed MAC algorithm supports the hierarchical primary-secondary user structure in the following sense.

Theorem 3: Let K_p be the number of primary users in the system. The value of K_p is unknown to the users as well as to the receiver. With the proposed MAC algorithm, the system should possess a unique equilibrium. Let channel contention measure at the equilibrium be denoted by q_v. On one hand, if K_p is small such that $q_{vp}^*(K_p) \geq \underline{q_v}$, then $q_v \geq \underline{q_v}$ must hold at the equilibrium. On the other hand, if K_p is large such that $q_{vp}^*(K_p) < \underline{q_v}$, then transmission probabilities of the secondary users should equal zero at the equilibrium.

Proof: According to the stochastic approximation framework presented in [15, Section 4.1], the system should have at least one equilibrium.

We first show that the equilibrium must be unique. Assume that this is not true. Let the system contain two equilibria, whose corresponding channel contention measures equal q_v and $\tilde{q}_v$, respectively. Without loss of generality, we assume that $q_v < \tilde{q}_v$. Assume that, at the first equilibrium corresponding to channel contention measure q_v, the number of users estimated by the primary users and by the secondary users equal respectively $\hat{K}_p$ and $\hat{K}_s$. At the other equilibrium corresponding to channel contention measure $\tilde{q}_v$, let the estimates equal $\tilde{K}_p$ and $\tilde{K}_s$, respectively. Consequently, we have

$$\begin{aligned} q_v &= q_{vp}^*(\hat{K}_p) = q_{vs}^*(\hat{K}_s), \\ \tilde{q}_v &= q_{vp}^*(\tilde{K}_p) = q_{vs}^*(\tilde{K}_s). \end{aligned} \tag{20}$$

Because $q_v < \tilde{q}_v$, due to the fact that $q_{vp}^*(\hat{K})$ and $q_{vs}^*(\hat{K})$ functions are non-increasing in $\hat{K}$ [15, Theorem 4.4], (20) implies that $\hat{K}_p \geq \tilde{K}_p$ and $\hat{K}_s \geq \tilde{K}_s$. This consequently implies that $p_p^*(\hat{K}_p) \leq p_p^*(\tilde{K}_p)$ and $p_s^*(\hat{K}_p) \leq p_s^*(\tilde{K}_p)$. However, if each user at the first equilibrium should transmit at a probability no higher than the corresponding probability at the other equilibrium, we must have $q_v \geq \tilde{q}_v$, which contradicts the assumption that $q_v < \tilde{q}_v$. Therefore, equilibrium of the system must be unique.

Let q_v be the channel contention measure at the unique equilibrium. Next, we prove the following statement, which is equivalent to the conclusion of the theorem. That is, if $q_v < \underline{q_v}$, we must have $q_{vp}^*(K_p) < \underline{q_v}$. Otherwise if $q_v \geq \underline{q_v}$, we must have $q_{vp}^*(K_p) \geq \underline{q_v}$.

According to the proposed MAC algorithm, if $q_v < \underline{q_v} \leq q_{vs}^*(\hat{K})$ for all $\hat{K}$, we must have $\hat{K}_s = \infty$ and all secondary users should have zero transmission probability at the equilibrium. Consequently, the system becomes equivalent to one with homogeneous (primary) users, as analyzed in [15]. According to [15, Theorem 4.5], we should have

$$q_v = q_v(p_p^*(K_p), K_p) = q_{vp}^*(K_p). \tag{21}$$

In other words, primary users should obtain a correct estimate of the number of users $\hat{K}_p = K_p$. This implies that $q_{vp}^*(K_p) = q_{vp}^*(\hat{K}_p) = q_v < \underline{q_v}$.

If $q_v \geq \underline{q_v}$, on the other hand, we have $q_v = q_{vp}^*(\hat{K}_p)$. In this case, $\hat{K}_s < \infty$, meaning that secondary users should transmit with a positive probability. Now assume that we force all secondary users to exit the system. This action should help increasing the value of q_v at the new equilibrium. We know that, without the secondary users, contention measure of the new system equilibrium should equal $q_v(p_p^*(K_p), K_p) = q_{vp}^*(K_p)$. Consequently, we must have

$$\underline{q_v} \leq q_v \leq q_v(p_p^*(K_p), K_p) = q_{vp}^*(K_p). \tag{22}$$

■

Example 1: Consider a random multiple access system over a classical collision channel. Assume that the virtual packet should be identical to a real packet, and therefore virtual packet reception should be regarded as successful if and only if all users idle in a time slot. In this case, channel contention measure q_v, which is the success probability of the virtual packet, should equal the idling probability of the collision channel.

Assume that primary users intend to maximize the symmetric throughput if there is no secondary user in the system. In other words, under the assumption that there are K primary users, and all users have the same transmission probability p, utility function of the primary users is given by $U_p(K, p) = p(1-p)^{K-1}$. According to [15, Section 4.2], we can choose the desired transmission probability function of the primary users as $p_p^*(\hat{K}) = \frac{1}{\max\{\hat{K},1\}+1.01}$. This implies that $K_{p\min} = 1$, $b_p = 1.01$ and $x_p^* = 1$. Consequently, the theoretical channel contention measure function for the primary users is given by $q_{vp}^*(\hat{K}) = \left(1 - \frac{1}{\hat{K}+1.01}\right)^{\hat{K}}$ for $\hat{K} \geq 1$.

Next, let $\underline{q_v}$ be set at $\underline{q_v} = e^{-0.85} = 0.427$. This gives $x_s^* = 0.85$ according to (17). Consequently, we can choose the desired transmission probability function for the secondary users as $p_s^*(\hat{K}) = \frac{0.85}{\max\{\hat{K},1\}+1.01}$, for $\hat{K} \geq 1$. This implies that $b_s = 1.01$ and $K_{s\min} = 1$. The function of theoretical channel contention measure for the secondary users is given by $q_{vs}^*(\hat{K}) = \left(1 - \frac{0.85}{\hat{K}+1.01}\right)^{\hat{K}}$ for $\hat{K} \geq 1$. Note that $q_{vs}^*(\hat{K}) \geq \underline{q_v} = e^{-0.85} = 0.427$.

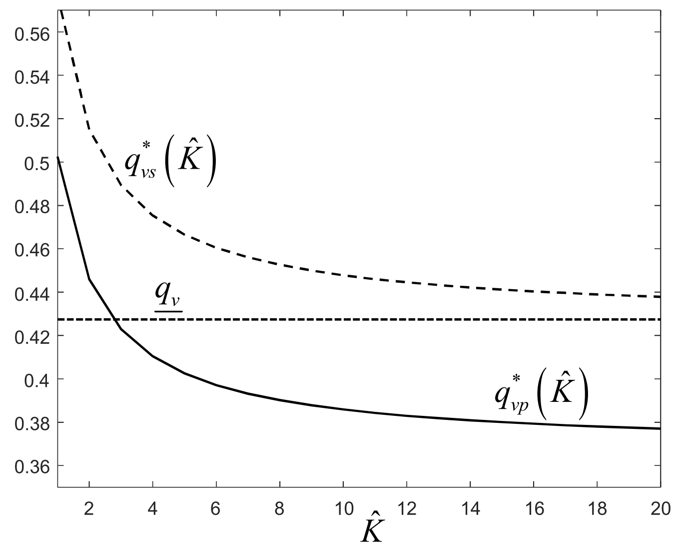


Fig. 1. Theoretical channel contention measure functions for primary and secondary users.

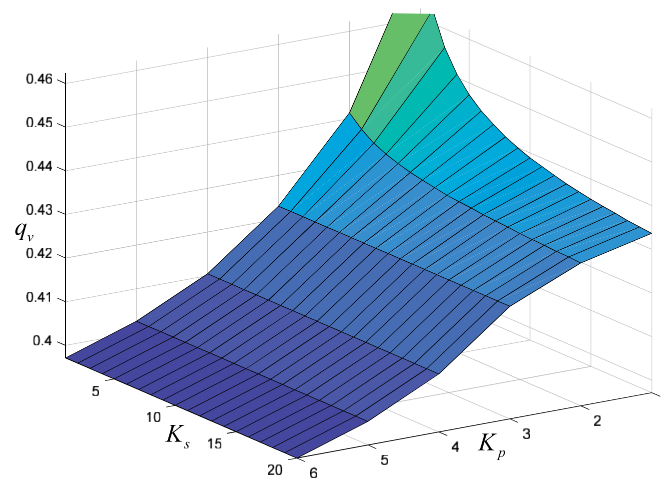


Fig. 2. Channel idling probability as a function of the numbers of primary and secondary users.

In Figure 1, we plotted the theoretical channel contention measure functions $q_{vp}^*(\hat{K})$ for primary users and $q_{vs}^*(\hat{K})$ for secondary users. It can be seen that, key idea of supporting the hierarchical user structure is to raise the tail of the $q_{vs}^*(\hat{K})$ function for the secondary users above $\underline{q_v}$, such that aggregated impact of the secondary users on the idling probability of the channel is well controlled no matter how many secondary users are competing for the channel.

In Figure 2, we plotted channel contention measure of the system at its unique equilibrium as a function of the number of primary users K_p and the number of secondary users K_s. The figure shows that, when the number of primary users is small $K_p \leq 2$, we have $q^*_{vp}(K_p) > \underline{q_v} = 0.427$. In this case, secondary users can access the channel. But the system keeps the channel idling probability above $\underline{q_v} = 0.427$ irrespective of the number of secondary users. When the number of primary users is large $K_p > 2$, on the other hand, we have $q^*_{vp}(K_p) < \underline{q_v}$. In this case transmission probabilities of the secondary users are kept at zero, and therefore q_v is not affected by the number of secondary users.

IV. Supporting Hierarchical Users With Multiple Transmission Options

In this section, we consider the general scenario when each user is equipped with M transmission options plus an idling option, where M is a positive integer. A user, for example user k, is associated with a transmission probability vector $\boldsymbol{p}_k = p_k \boldsymbol{d}_k$, where p_k is the transmission probability and $\boldsymbol{d}_k$ is the transmission direction vector. Under the assumption that all users have the same transmission direction vector $\boldsymbol{d}$, the link layer multiple access channel is modeled using two sets of parameter functions, namely the "real channel parameter function set" $\{C_{rij}(\boldsymbol{d})\}$ and the "virtual channel parameter function set" $\{C_{vj}(\boldsymbol{d})\}$ [15, Section 4.3].

Let the system have K_p primary users and K_s secondary users, with $K = K_p + K_s$. If all users have the same transmission probability vector $\boldsymbol{p} = p\boldsymbol{d}$, "channel contention measure" q_v can be calculated by

$$q_v(p\boldsymbol{d}, K) = \sum_{j=0}^{K} \binom{K}{j} p^j (1-p)^{K-j} C_{vj}(\boldsymbol{d}). \tag{23}$$

We again assume that, upon receiving the feedback of q_v, each primary (secondary) user should obtain an estimated number of users, denoted by $\hat{K}$ based on the assumption that the system should only contain primary (secondary) users. Each primary (secondary) user should design two key functions, denoted by $\boldsymbol{p}^*_p(\hat{K})$ and $q^*_{vp}(\hat{K})$ ($\boldsymbol{p}^*_s(\hat{K})$ and $q^*_{vs}(\hat{K})$). $\boldsymbol{p}^*_p(\hat{K})$ ($\boldsymbol{p}^*_s(\hat{K})$) represents the theoretical transmission probability vector of each user if the number of users equal $\hat{K}$. $q^*_{vp}(\hat{K})$ ($q^*_{vs}(\hat{K})$), on the other hand, represents the "theoretical channel contention measure" when the number of users equals $\hat{K}$ and all users have the same transmission probability vector $\boldsymbol{p}^*_p(\hat{K})$ ($\boldsymbol{p}^*_s(\hat{K})$). With the two key functions, the distributed MAC algorithm remains the same as introduced in Section III, with the only revision that transmission probabilities p_{k_p} and p_{k_s} in the algorithm of Section III should be replaced by transmission probability vectors $\boldsymbol{p}_{k_p}$ and $\boldsymbol{p}_{k_s}$, respectively.

However, as explained in [15, Section 4.3], the design of the two key functions in this case is much more complicated than that of the single transmission option scenario. The key challenge is that, because $\boldsymbol{p}^*_p(\hat{K}) = p^*_p(\hat{K})\boldsymbol{d}^*_p(\hat{K})$ ($\boldsymbol{p}^*_s(\hat{K}) = p^*_s(\hat{K})\boldsymbol{d}^*_s(\hat{K})$)) often involves different transmission direction vectors $\boldsymbol{d}^*_p(\hat{K})$ ($\boldsymbol{d}^*_s(\hat{K})$) for different $\hat{K}$ values, it becomes mathematically difficult to argue for any monotonicity property in $\hat{K}$. To overcome this challenge, [15, Section 4.3] proposed a search-assisted design approach. The proposed approach achieves close to optimal performance by manually designing $\boldsymbol{p}^*_p(\hat{K})$ and $q^*_{vp}(\hat{K})$ ($\boldsymbol{p}^*_s(\hat{K})$ and $q^*_{vs}(\hat{K})$) for selected $\hat{K}$ values, and then uses an automatic interpolation approach to complete the two functions to achieve the desired monotonicity property for $q^*_{vp}(\hat{K})$ ($q^*_{vs}(\hat{K})$) in $\hat{K}$ [15, Section 4.3]. For the sake of completeness, we will briefly outline the design approach in this paper. While reading the full design approach of [15, Section 4.3] is recommended, skipping the details won't compromise understandings on the core idea of supporting the hierarchical user structure.

Let us first focus on a primary user. We assume that primary users intend to maximize a utility of $U_p(\hat{K}, p_p\boldsymbol{d}_p, \{C_{rij}(\boldsymbol{d}_p)\})$, under the assumption that the system contains $\hat{K}$ homogeneous primary users and all users have the same transmission probability vector $\boldsymbol{p}_p = p_p\boldsymbol{d}_p$. A primary user should first identify two integers $\underline{K_p}$ and $\overline{K_p}$, with $\underline{K_p} \leq \overline{K_p}$. Define $\{\hat{K} : \hat{K} \leq \underline{K_p}\}$ as the "head regime" and define $\{\hat{K} : \hat{K} \geq \overline{K_p}\}$ as the "tail regime". Consider the head regime first. A primary user should find the optimal transmission direction vector $\underline{\boldsymbol{d}^*_p}$ defined as $\underline{\boldsymbol{d}^*_p} = \arg\max_{\boldsymbol{d}_p} \max_{p_p} U_p(\underline{K_p}, p_p\boldsymbol{d}_p, \{C_{rij}(\boldsymbol{d}_p)\})$, and then assume that transmission direction vectors should be fixed at $\underline{\boldsymbol{d}^*_p}$ within the head regime, i.e., $\boldsymbol{p}^*_p(\hat{K}) = p^*_p(\hat{K})\underline{\boldsymbol{d}^*_p}$, for $\hat{K} \leq \underline{K_p}$. Consequently, the design problem of $\boldsymbol{p}^*_p(\hat{K})$ and $q^*_{vp}(\hat{K})$ for $\hat{K} \leq \underline{K_p}$ becomes equivalent to the scenario when each user only has a single transmission option [15, Section 4.3]. This means that $p^*_p(\hat{K})$ can be set at $p^*_p(\hat{K}) = \frac{x^*_p}{\max\{\hat{K}, \hat{K}_{p\min}\} + b_p}$ with parameters x^*_p, $\hat{K}_{p\min}$, b_p being chosen according to the guideline provided in Section III and in [15, Section 4.2]. $q^*_{vp}(\hat{K})$ for $\hat{K} \leq \underline{K_p}$ can then be determined according to (14) and (15). Similarly, for the tail regime, a primary user should first find the optimal transmission direction vector $\overline{\boldsymbol{d}^*_p} = \arg\max_{\boldsymbol{d}_p} \max_{p_p} U_p(\overline{K_p}, p_p\boldsymbol{d}_p, \{C_{rij}(\boldsymbol{d}_p)\})$, and then design $\boldsymbol{p}^*_p(\hat{K}) = p^*_p(\hat{K})\overline{\boldsymbol{d}^*_p}$ and $q^*_{vp}(\hat{K})$ for $\hat{K} \geq \overline{K_p}$ accordingly under the assumption that transmission direction vectors should be fixed at $\overline{\boldsymbol{d}^*_p}$.

For $\hat{K}$ between the head and the tail regimes, a primary user should first manually design $\boldsymbol{p}^*_p(\hat{K})$ and $q^*_{vp}(\hat{K})$ for several selected $\hat{K}$ values, termed "pinpoints". $\boldsymbol{p}^*_p(\hat{K})$ and $q^*_{vp}(\hat{K})$ functions can then be completed using the interpolation approach introduced in [15, Section 4.3]. Note that, $\underline{K_p}$, $\overline{K_p}$, the pinpoints and the corresponding function values must be carefully designed to satisfy a set of constraints articulated in [15, Section 4.3]. These constraints intend to achieve two key design objectives. First, $\boldsymbol{p}^*_p(\hat{K})$ should be close to optimal in terms of maximizing the chosen utility of $U_p(\hat{K}, p_p\boldsymbol{d}_p, \{C_{rij}(\boldsymbol{d}_p)\})$. Second, $q^*_{vp}(\hat{K})$ should be monotonically non-increasing in $\hat{K}$ and should be strictly decreasing in $\hat{K}$ for $\hat{K} \geq \hat{K}_{p\min}$ [15, Section 4.3].

Next, consider a secondary user. The design procedure of the two key functions $\boldsymbol{p}^*_s(\hat{K})$ and $q^*_{vs}(\hat{K})$ for a secondary user is essentially similar to that of a primary user, except that a secondary user needs to meet the additional requirement of enforcing the hierarchical user structure. Similar to the single transmission option scenario discussed in Section III, the core idea of establishing the hierarchical user structure is to raise the tail of the $q^*_{vs}(\hat{K})$ function above a pre-determined threshold $\underline{q_v}$. The value of $\underline{q_v}$ should be determined according to the quality of service requirement of the primary users. Let $U_s(\hat{K}, p_s\boldsymbol{d}_s, \{C_{rij}(\boldsymbol{d}_s)\})$ be the utility function chosen by the secondary users, under the assumption that the system contains $\hat{K}$ homogeneous secondary users and all users have the same transmission probability vector $\boldsymbol{p}_s = p_s\boldsymbol{d}_s$. A secondary user should identify two integers $\underline{K_s}$ and $\overline{K_s}$ with $\underline{K_s} \le \overline{K_s}$, and define the "head regime" $\{\hat{K} : \hat{K} \le \underline{K_s}\}$ as well as the "tail regime" $\{\hat{K} : \hat{K} \ge \overline{K_s}\}$. Consider the tail regime first. A secondary user should find the optimal transmission direction vector $\overline{\boldsymbol{d}_s^*}$ and assume that transmission direction vectors should be fixed at $\overline{\boldsymbol{d}_s^*}$ within the tail regime, i.e., $\boldsymbol{p}_s^*(\hat{K}) = p_s^*(\hat{K})\overline{\boldsymbol{d}_s^*}$. Let $\overline{x_s^*}$ be determined by

$$\lim_{\hat{K}\to\infty} q_v\left(\frac{\overline{x_s^*}}{\hat{K}}\overline{\boldsymbol{d}_s^*}, \hat{K}\right) = \underline{q_v}, \tag{24}$$

where $q_v(.)$ can be further calculated using (23). Then, a secondary user can set $p_s^*(\hat{K})$ for the tail regime at $p_s^*(\hat{K}) = \frac{\overline{x_s^*}}{\max\{\hat{K}, \overline{\hat{K}_{s\min}}\}+\overline{b_s}}$, where parameters $\overline{\hat{K}_{s\min}}$, $\overline{b_s}$ should be chosen according to the guideline provided in Section III and in [15, Section 4.2].

The rest of the design of $\boldsymbol{p}_s^*(\hat{K})$ and $q_{vs}^*(\hat{K})$ can proceed similar to that of a primary user. Note that, while the constraint of maintaining the hierarchical structure appears to impact only the design in the tail regime, because the design needs to ensure that $q_{vs}^*(\hat{K})$ should be monotonically non-increasing in $\hat{K}$, the constraint is likely to spill over to impact the design of the pinpoints for $\hat{K}$ with $\underline{K_s} \le \hat{K} \le \overline{K_s}$, and may also impact the design in the head regime for $\hat{K} \le \underline{K_s}$.

With the design illustrated above, the distributed MAC algorithm supports the hierarchical user structure in the following sense.

Theorem 4: Let K_p be the number of primary users in the system. The value of K_p is unknown to the users as well as to the receiver. Let q_v be the channel contention measure at an equilibrium of the system. With the proposed MAC algorithm, on one hand, if K_p is small such that $q_{vp}^*(K_p) \ge \underline{q_v}$, then $q_v \ge \underline{q_v}$ must hold. On the other hand, if K_p is large such that $q_{vp}^*(K_p) < \underline{q_v}$, then transmission probabilities of the secondary users at any equilibrium of the system should equal zero.

The proof of Theorem 4 can be obtained from the proof of Theorem 3 with only minor revisions.

Note that when users are equipped with multiple transmission options, we can no longer prove uniqueness of the system equilibrium, even though we believe that the equilibrium should indeed be unique under mild conditions. The challenge comes from the fact that, due to generality of the system model, it is mathematically difficult to compare channel contention measures when users change their transmission direction vectors. Also note that, in the design approach presented in Section III, we raise the tail of $q_{vs}^*(\hat{K})$ by choosing an appropriate value of x_s^*, which affects the value of $p_s^*(\hat{K})$ for all $\hat{K}$ values. In the design approach presented in this section, however, it is possible to raise the tail of $q_{vs}^*(\hat{K})$ by limiting the impact of such a constraint mainly toward the tail regime. While the latter approach gives more flexibility in designing the $\boldsymbol{p}_s^*(\hat{K})$ function, it requires the search-assisted approach as introduced in [15, Section 4.3], which does not yield closed-form expressions for the $p_p^*(\hat{K})$ and $p_s^*(\hat{K})$ functions.

Example 2: This example is revised from [15, Example 4.4]. Consider a time-slotted multiple access network over a multi-packet reception channel. Each user is equipped with two transmission options where the first option is a high-rate option and the second option is a low-rate option, respectively. If all packets are encoded using the low-rate option, then the channel can support the parallel transmissions of no more than 12 packets. We assume that one packet from the high-rate option is equivalent to the combination of 4 low-rate packets. Therefore, the channel can support the parallel transmissions of n_1 high-rate packets plus n_2 low-rate packets if and only if $\frac{1}{3}n_1 + \frac{1}{12}n_2 \le 1$. We design the virtual packet to be equivalent to a high rate real packet. Consequently, the two sets of channel parameter functions $\{C_{rij}(\boldsymbol{d})\}$ and $\{C_{vj}(\boldsymbol{d})\}$ can be derived accordingly.

First, consider a primary user. We assume that primary users intend to maximize the sum system throughput. That is, under the assumption that the system has K homogeneous primary users and all users have the same transmission probability vector $\boldsymbol{p} = [p_1, p_2]^T = p[d_1, d_2]^T$, utility function of the primary users is given by

$$\begin{aligned} &U_p(K, \boldsymbol{p}, \{C_{rij}(\boldsymbol{d})\}) \\ &= K\sum_{i=1}^{2} d_i r_i \times \sum_{j=0}^{K-1} \binom{K-1}{j} p^{j+1}(1-p)^{K-1-j} C_{rij}(\boldsymbol{d}), \end{aligned} \tag{25}$$

where $r_1 = 4$, $r_2 = 1$ are the rate parameters of the two options.

The desired transmission probability vector function $\boldsymbol{p}_p^*(\hat{K})$ for the primary users is designed as follows. We define $\{\hat{K} : \hat{K} \le 4\}$ as the head regime, and set $\boldsymbol{p}_p^*(\hat{K})$ for the head regime at $\boldsymbol{p}_p^*(\hat{K}) = \frac{2.27}{\max\{\hat{K},2\}+1.01}[1,0]^T$. This implies that $\underline{K_p} = 4$, $\underline{\boldsymbol{d}_p^*} = [1,0]^T$, $\underline{K_{p\min}} = 2$, $\underline{b_p} = 1.01$ and $\underline{x_p^*} = 2.27$, which is determined using the following formula.

$$\underline{x_p^*} = \arg\max_{x} \lim_{\hat{K}\to\infty} U_p\left(\hat{K}, \frac{x}{\hat{K}}[1,0]^T, \{C_{rij}([1,0]^T)\}\right). \tag{26}$$

We define $\{\hat{K} : \hat{K} \ge 10\}$ as the tail regime, and set $\boldsymbol{p}_p^*(\hat{K})$ for the tail regime at $\boldsymbol{p}_p^*(\hat{K}) = \frac{8.82}{\hat{K}+1.01}[0,1]^T$. This implies that $\overline{K_p} = 10$, $\overline{\boldsymbol{d}_p^*} = [0,1]^T$, $\overline{K_{p\min}} = 8$, $\overline{b_p} = 1.01$ and $\overline{x_p^*} = 8.82$, which is determined using the following formula.

$$\overline{x_p^*} = \arg\max_{x} \lim_{\hat{K}\to\infty} U_p\left(\hat{K}, \frac{x}{\hat{K}}[0,1]^T, \{C_{rij}([0,1]^T)\}\right). \tag{27}$$

The theoretical channel contention measure function $q_{vp}^*(\hat{K})$ for the head and the tail regimes can be calculated accordingly.

For $4 \le \hat{K} \le 10$, in addition to $\hat{K} = 4$ and $\hat{K} = 10$, we choose two extra pinpoints at $\hat{K} = 5$ and $\hat{K} = 6$. We set transmission direction vectors $\boldsymbol{d}_p^*(5)$ and $\boldsymbol{d}_p^*(6)$ at the same direction vectors corresponding to the probability vectors that

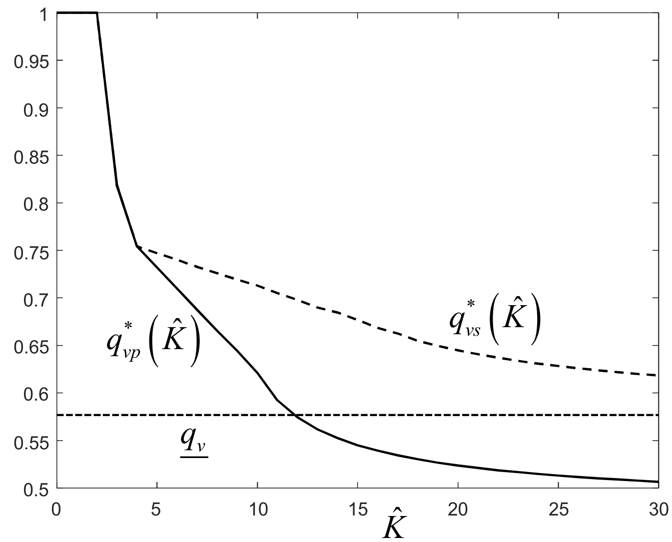


Fig. 3. Theoretical channel contention measure functions for primary and secondary users.

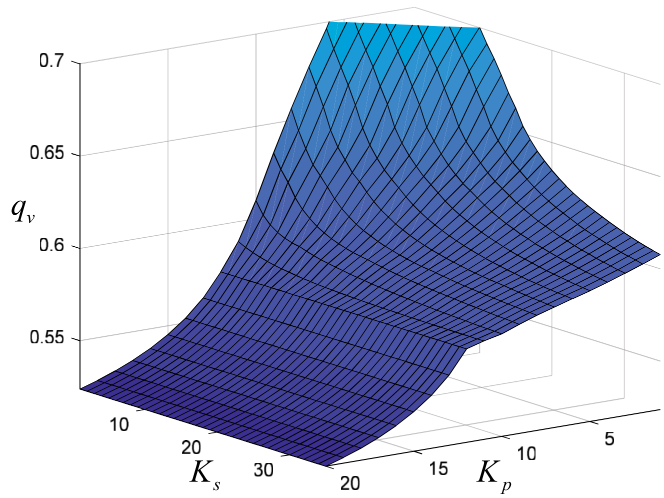


Fig. 4. Channel contention measure at the equilibrium as a function of the number of primary users and the number of secondary users.

maximize the utility (25) for $\hat{K} = 5$ and $\hat{K} = 6$, respectively. With $\boldsymbol{d}_p^*(5)$, $\boldsymbol{d}_p^*(6)$, and $\boldsymbol{d}_p^*(4) = [1, 0]^T$, $\boldsymbol{d}_p^*(10) = [0, 1]^T$, we then set $\boldsymbol{d}_p^*(\hat{K})$ for $4 \le \hat{K} \le 10$ such that $\boldsymbol{d}_p^*(\hat{K})$ transit linearly in $\hat{K}$ between the neighboring pinpoints. Next, we choose $p_p^*(\hat{K})$ for $4 \le \hat{K} \le 10$ such that the resulting $q_{vp}^*(\hat{K})$ function is linear in $\hat{K}$ for $4 \le \hat{K} \le 10$.

Now consider a secondary user. Assume that the hierarchical user structure requires $q_{vs}^*(\hat{K}) \ge \underline{q_v} = e^{-0.55} = 0.577$. We define $\{\hat{K} : \hat{K} \ge 18\}$ as the tail regime for secondary users, i.e., $\overline{K_s} = 18$. $\boldsymbol{p}_s^*(\hat{K})$ for the tail regime is set at $\boldsymbol{p}_s^*(\hat{K}) = \frac{8.11}{\hat{K}+1.01}[0, 1]^T$. This implies that $\overline{\boldsymbol{d}_s^*} = [0, 1]^T$, $\overline{K_{s\,\min}} = 8$, $\overline{b_s} = 1.01$ and $\overline{x_s^*} = 8.11$, which is determined by solving equation (24) with $\underline{q_v} = e^{-0.55} = 0.577$.

Because the value of $q_{vs}^*(\overline{K_s})$ is reasonably low, we choose to set the head regime design of a secondary user the same as that of a primary user. In other words, $\underline{K_s} = 4$, $\boldsymbol{p}_s^*(\hat{K}) = \boldsymbol{p}_p^*(\hat{K})$ and $q_{vs}^*(\hat{K}) = q_{vp}^*(\hat{K})$ for $\hat{K} \le 4$.

For $4 \le \hat{K} \le 18$, we simply set $\boldsymbol{d}_s^*(\hat{K}) = \boldsymbol{d}_p^*(\hat{K})$, and then choose $p_s^*(\hat{K})$ such that the resulting $q_{vs}^*(\hat{K})$ function is linear in $\hat{K}$ for $4 \le \hat{K} \le 18$.

Figure 3 shows the theoretical channel contention measure functions $q_{vp}^*(\hat{K})$ and $q_{vs}^*(\hat{K})$, respectively. In this example, because we are able to keep the same design for both primary and secondary users for $\hat{K} \le 4$, they are treated equally if the total number of users is no larger than 4. Whether such a design is feasible or not depends on the value of $\underline{q_v}$. Note that the design is enabled due to the flexibility of the search-assisted design approach.

In this example, although we are not able to prove that equilibrium of the system should be unique, it is indeed the case according to numerical search. In Figure 4, channel contention measure q_v at the equilibrium is plotted as a function of the number of primary users K_p and the number of secondary users K_s. It can be seen that, when the number of primary users is small $K_p < 12$, we have $q_{vp}^*(K_p) > \underline{q_v} = e^{-0.55} = 0.577$. In this case, the system allows secondary users to access the channel but keeps the channel contention measure q_v above $\underline{q_v} = e^{-0.55} = 0.577$ irrespective of the number of secondary users. When the number of primary users is large $K_p \ge 12$, on the other hand, we have $q_{vp}^*(K_p) < \underline{q_v}$. In this case transmission probabilities of the secondary users are kept at zero, and therefore q_v is not affected by the number of secondary users.

Next, we assume that the system has 6 primary users and 10 secondary users initially. Transmission probabilities of the users are initialized at $[0, 0]^T$. In each time slot, a user randomly determines whether to transmit a packet or not, and if the answer is positive, which option should be used. The receiver uses an exponential moving average approach to measure q_v. More specifically, q_v is initialized at $q_v = 1$. In each time slot, an indicator variable $I_v \in \{0, 1\}$ is used to represent the success/failure status of the virtual packet reception. q_v is then updated as $q_v = (1 - \frac{1}{300})q_v + \frac{1}{300}I_v$, and is fed back to the transmitters at the end of each time slot. With the updated q_v, each user adapts its transmission probability vector according to the MAC algorithm proposed in Section IV with a constant step size of $\alpha = 0.05$.

We assume that the system experiences three stages. At the beginning in Stage one, the system has 6 primary users. The system enters Stage two after the 3000th time slot, when 6 more primary users enter the system with their transmission probability vectors initialized at $[0, 0]^T$. After the 6000th time slot, the system enters Stage three when 9 primary users exit the system. Throughout the three stages, the number of secondary users is kept at 10. Convergence behavior in actual channel contention measure q_v is illustrated in Figure 5 together with the theoretical q_v at the corresponding equilibria of the three stages. The figure demonstrates that the system can quickly adapt to changes in the number of users and keep the channel contention at the desired level. In Figure 6, we also illustrated entries of the transmission probability vector targets calculated by the primary and the secondary users. Note that values of the simulated variables presented in Figures 5 and 6 are calculated using the same exponential averaging approach

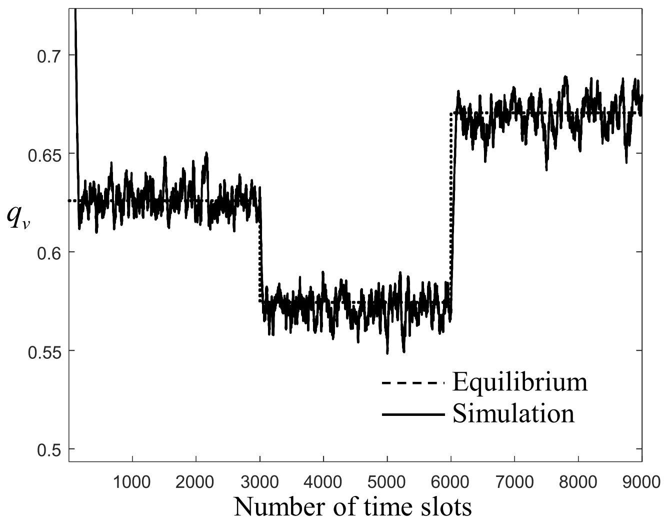


Fig. 5. Channel contention measure of the system through three stages.

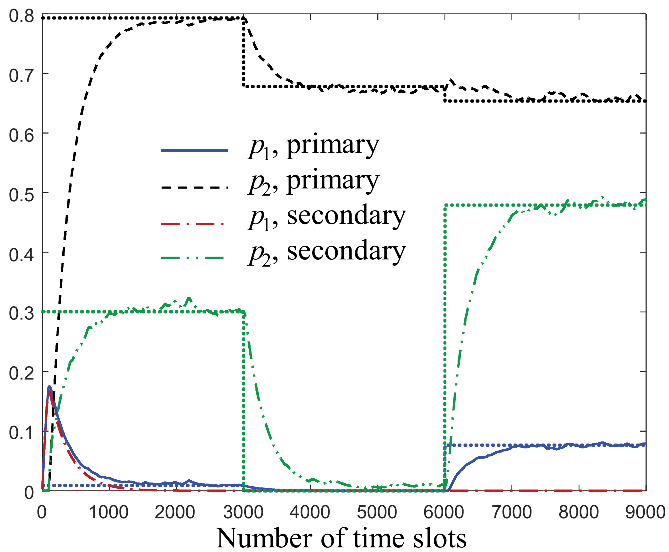


Fig. 6. Transmission probabilities of primary and secondary users through three stages. Dashed lines represent the corresponding values at the equilibrium.

explained above. It can be seen that the system is reasonably responsive to changes in the number of users and can quickly lead transmission probability vectors of both primary users and secondary users to their corresponding theoretical equilibrium values. The hierarchical user structure can be seen clearly in the sense that secondary users always transmit with a low rate option at a relatively low probability.

V. Further Discussion

In the DSA literature, hierarchical channel sharing approaches are categorized into "overlay" and "underlay" schemes [21], [22]. In an overlay scheme, secondary users can access the channel with significant transmission power but only when primary users are not present. In an underlay scheme, secondary users can access the channel under the constraint that their aggregated interference should be maintained below a pre-determined level. The distributed MAC algorithms introduced in Sections III and IV of this paper can be viewed as an underlay scheme for random multiple access systems where aggregated "interference" of the users is evaluated using the chosen channel contention measure. While most of the underlay schemes assume a pre-determined quality of service threshold, such as a pre-determined $\underline{q_v}$ assumed in this paper, the threshold can be made adaptive according to, for example, broadcasted requests from the primary users.

The distributed MAC algorithm can be easily extended to support user groups with more than two hierarchical levels by raising tails of the theoretical channel contention measure functions of different user groups to different pre-determined thresholds. It can also be extended to further incorporate a "soft" priority structure between users groups in addition to the already established "hard" hierarchical user structure. Take the primary-secondary user structure discussed in this paper for example. The "hard" hierarchical user structure is established by raising the tail of the $q^*_{vs}(\hat{K})$ function to $q^*_{vs}(\infty) = \underline{q_v}$. In addition, when the number of users is small, the system can further control the relative priority of the user groups by imposing a required distance (horizontally at the same q_v level) between the two functions of $q^*_{vp}(\hat{K})$ and $q^*_{vs}(\hat{K})$. More specifically, let $\hat{K}_{q^*_{vp}}(q_v)$ and $\hat{K}_{q^*_{vs}}(q_v)$ be the inverse functions of $q^*_{vp}(\hat{K})$ and $q^*_{vs}(\hat{K})$, respectively. We can require the design of the two functions to satisfy $\hat{K}_{q^*_{vs}}(q_v) \geq A_0\hat{K}_{q^*_{vp}}(q_v)$ for all $q_v > \underline{q_v}$, where $A_0 \geq 1$ is a pre-determined constant. Such a requirement can guarantee that the estimated number of users for a secondary user should always be significantly larger than that of a primary user. Consequently, a "soft" priority structure is established in the sense that the derived target transmission probability of a secondary user should always be significantly smaller than that of a primary user. However, further investigation on incorporating a "soft" priority user structure is beyond the scope of this paper.

VI. Conclusion

We proposed a distributed MAC framework to support hierarchical user groups in random multiple access systems. The MAC algorithms do not require direct message exchange among users. Users do not need to know the numbers of primary and secondary users in the system. Users also do not need the capability of identifying whether a transmitted packet should belong to a primary user or to a secondary user. The proposed MAC algorithm adapts the transmission scheme of each user by comparing the actual channel contention measure to a theoretical channel contention measure function. With the simple idea of raising the tail of the theoretical channel contention measure function for the secondary users to a pre-determined threshold, aggregated impact of the secondary users on contention level of the channel is well controlled no matter how many secondary users compete for the channel. We extended the proposed MAC algorithm to a system where each user is equipped with multiple transmission options. Simulation results showed that the proposed MAC algorithm can maintain the hierarchical user structure and can also be reasonably responsive to a dynamic environment with users joining/exiting the system.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard 802.11-1999, IEEE Computer Society LAN Man Standards Committee, New York, NY, USA, 1999.

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

[3] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1992.

[4] S. Ghez, S. Verdu, and S. C. Schwartz, "Stability properties of slotted aloha with multipacket reception capability," *IEEE Trans. Autom. Control*, vol. 33, no. 7, pp. 640–649, Jul. 1988.

[5] S. Ghez, S. Verdu, and S. C. Schwartz, "Optimal decentralized control in the random access multipacket channel," *IEEE Trans. Autom. Control*, vol. 34, no. 11, pp. 1153–1163, Nov. 1989.

[6] H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*. New York, NY, USA: Springer-Verlag, 1997.

[7] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, 2nd ed. San Diego, CA, USA: Academic, 1975.

[8] V. Borkar and S. Meyn, "The O.D.E method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, pp. 447–469, Jan. 2000.

[9] B. Hajek, "Stochastic approximation methods for decentralized control of multiaccess communications," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 2, pp. 176–184, Mar. 1985.

[10] B. Bensaou, Y. Wang, and C. Chung Ko, "Fair medium access in 802.11 based wireless ad-hoc networks," in *Proc. 1st Annu. Workshop Mobile Ad Hoc Netw. Comput. (MobiHOC)*, 2000, pp. 99–106.

[11] T. Nandagopal, T.-E. Kim, X. Gao, and V. Bharghavan, "Achieving MAC layer fairness in wireless packet networks," in *Proc. 6th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2000, pp. 87–98.

[12] G. D. Celik, G. Zussman, W. F. Khan, and E. Modiano, "MAC for networks with multipacket reception capability and spatially distributed nodes," in *Proc. IEEE 27th Conf. Comput. Commun. (INFOCOM)*, Apr. 2008, pp. 1436–1444.

[13] J. Luo and A. Ephremides, "A new approach to random access: Reliable communication and reliable collision detection," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 989–1002, Feb. 2012.

[14] J. Luo, "A generalized channel coding theory for distributed communication," *IEEE Trans. Commun.*, vol. 63, pp. 1043–1056, Apr. 2015.

[15] Y. Tang, F. Heydaryan, and J. Luo, "Distributed coding in a multiple access environment," *Found. Trends Netw.*, vol. 12, no. 4, pp. 260–412, 2018.

[16] Y. Tang, F. Heydaryan, and J. Luo, "Distributed multiple access with a general link layer channel," *Ad Hoc Netw.*, vol. 85, pp. 120–130, Mar. 2019.

[17] I. Stavrakakis and D. Kazakos, "A multiuser random-access communication system for users with different priorities," *IEEE Trans. Commun.*, vol. 39, no. 11, pp. 1538–1541, Nov. 1991.

[18] H. Tang, M. Lei, X. Wang, and C. Sun, "The priority backoff algorithm and PUSCH power control for random access in TD-LTE trunking system," in *Proc. 2nd Int. Conf. Bus. Comput. Global Informatization*, Shanghai, China, Oct. 2012, pp. 580–583.

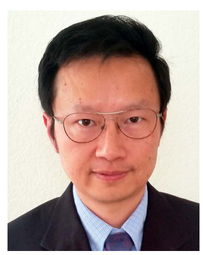
[19] L. Guan, B. Yan, Z. Guo, and Y. Gong, "Priority-based random access control mechanism for M2M communications," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Oct. 2016, pp. 2313–2317.

[20] *IEEE Standard for Information Technology–Local and Metropolitan Area Networks–Specific Requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications–Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Standard 802.11-2005, IEEE Computer Society LAN Man Standards Committee, New York, NY, USA, 2005.

[21] A. De Domenico, E. Calvanese Strinati, and M.-G. Di Benedetto, "A survey on MAC strategies for cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 1, pp. 21–44, 1st Quart., 2012.

[22] M. G. Khoshkholgh, K. Navaie, and H. Yanikomeroglu, "Access strategies for spectrum sharing in fading environment: Overlay, underlay, and mixed," *IEEE Trans. Mobile Comput.*, vol. 9, no. 12, pp. 1780–1793, Dec. 2010.

Faeze Heydaryan received the B.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 2010, the M.Sc. degree in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2012, and the Ph.D. degree in electrical and computer engineering from Colorado State University in 2020. Her main research interests include wireless networking and information theory.



Jie Luo (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Fudan University, Shanghai, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Connecticut in 2002. From September 2002 to August 2006, he was a Research Associate with the Institute for Systems Research (ISR), University of Maryland, College Park, MD, USA. In 2006, he joined Colorado State University, Fort Collins, CO, USA, where he is currently a Professor with the Department of Electrical and Computer Engineering. His research focuses on cross-layer design of wireless communication networks, with an emphasis on the bottom several layers. His general research interests include wireless communications, wireless networks, information theory, and signal processing.