Age-of-Information Oriented Scheduling for Multichannel IoT Systems With Correlated Sources

Jingwen Tong, *Student Member, IEEE*, Liqun Fu, *Senior Member, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*— Age-of-information (AoI) based minimization problems have been widely considered in Internet-of-Things (IoT) networks with the settings of multi-source single-channel systems and multi-source multi-channel systems. Most existing works are limited to either the case of identical multi-channel or independent sources. In this paper, we study this problem under the identical and non-identical multi-channel, as well as the correlated sources setting. This correlation defines the case when updating a source's AoI; others correlated to this one will also reveal partial information. To tackle this AoI-based minimization problem, we formulate it as a correlated restless multi-armed bandit (CRMAB) problem. By decoupling the CRMAB problem into N independent single-armed bandit problems, we derive the closed-form expressions of the generalized Whittle index (GWI) and the generalized partial Whittle index (GPWI) under the identical channel and the non-identical channel settings, respectively. Then, we put forth the GWI-based and GPWI-based scheduling policies to solve this AoI-based minimization problem. In addition, we provide two lower numerical performance bounds for the proposed policies by solving the relaxed Lagrange problem of the decoupled CRMAB. Numerical results show that the proposed policies can achieve these lower bounds and outperform the state-of-the-art scheduling policies. Compared with the case of independent sources, the performance of the proposed policies in the case of correlated sources improves significantly, especially in high-density networks.**

***Index Terms*— Age-of-information (AoI), correlated sources, correlated restless multi-armed bandit (CRMAB), generalized Whittle index (GWI), generalized partial Whittle index (GPWI).**

I. Introduction

INTERNET-OF-THINGS (IoT), which benefits from the wide deployment of next-generation wireless networks, plays an increasing role in daily life and industry such as e-health, smart home, driving, and monitoring [1]. An IoT network typically consists of three parts: IoT devices, transmission network, and base station (BS) [2], [3]. The IoT devices, also known as sources, are often deployed artificially or randomly to perceive the physical characteristic of the environment, such as temperature, humidity, and pollution level. By exploiting the transmission network, the perceived samples are sent to the BS for information fusion, where they are processed to extract meaningful information. The accuracy of such extracted information depends on the freshness of the perceived samples at the BS, playing a critical role in the BS's decision-making. Therefore, a fundamental problem that arises in IoT networks is how to deal with the time-sensitive information and to ensure its freshness [4].

Recently, age-of-information (AoI), which is defined as the time elapsed since the generation of the latest packet delivered to the destination, has gained much attention in the literature [5]–[7]. Compared with conventional performance metrics, such as delay and throughput, AoI provides a new perspective to quantify the freshness or accuracy of the samples from a remote system. The focus of this work is to minimize the average network-level AoI by scheduling the status updates of these IoT devices.

AoI-based minimization problems are widely considered in IoT networks with different system settings. Refs. [8]–[11] study this problem with the setting of multi-source and single-channel in IoT networks, where several sources (or IoT devices) send their samples to the BS through a shared channel. Meanwhile, refs. [12]–[18] investigate this problem with the setting of multi-source and multi-channel, where several IoT devices report their state status to the BS using multiple channels. However, these works assume that different channels are stochastically identical. Recently, the non-identical (or heterogeneous) multi-channel setting is considered in [19], [20]. To address these AoI-based minimization problems, the above works propose various scheduling policies, such as the round robbin policy (i.e., the greedy or myopic policy) [8], max-weight policy [9], MDP-based threshold policy [10], [11], Lyapunov-based virtual queue policy [13], [14], stationary randomized policy [16], and Whittle index (WI)-based scheduling policy [16]–[19].

However, these works assume that different sources are independent, ignoring their correlations. In practice, samples of different sources are usually relevant due to the spatial and temporal correlation of the perceived physical processes [21]–[27]. For example, Fig. 1 shows a video monitoring system where several cameras (or IoT devices) are

Manuscript received 31 August 2021; revised 6 February 2022 and 12 May 2022; accepted 23 May 2022. Date of publication 8 June 2022; date of current version 11 November 2022. The work of Liqun Fu was supported by the National Natural Science Foundation of China under Grant 61771017. The work of Zhu Han was supported by the U.S. National Science Foundation under Grant CNS-2107216 and Grant CNS-2128368. The associate editor coordinating the review of this article and approving it for publication was H. S. Dhillon. *(Corresponding author: Liqun Fu.)*

Jingwen Tong and Liqun Fu are with the Department of Information and Communication Engineering, Xiamen University, Xiamen 361005, China (e-mail: tongjingwen@stu.xmu.edu.cn; liqun@xmu.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: hanzhu22@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3179305.

Digital Object Identifier 10.1109/TWC.2022.3179305

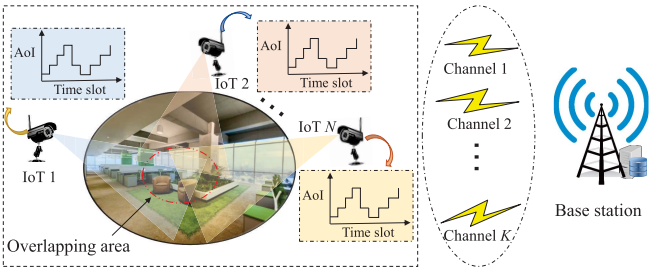


Fig. 1. A video monitoring system with multi-channel correlated sources.

deployed to monitor a specific area. There are some overlapped areas among these cameras since there are deployed artificially or randomly. Therefore, when updating a device's status, others overlapping with this one will also reveal partial monitoring information. In other words, the freshness (or AoI) of the information of the overlapped devices will also reduce. Another example is the forest fire monitoring system, where several sensors are deployed to monitor the fire in an area. The BS can infer a fire situation from some nearby sensors' monitoring area based on the status of the scheduled sensors. Thus, when updating a sensor's status, its neighbors will also reveal a partial fire situation. As a result, the BS can reduce the frequency of status updates by exploiting the correlations among IoT devices.

In this paper, we consider the AoI-based minimization problem in the IoT networks under the setting of stochastically identical and non-identical multi-channel, as well as the correlated sources. We first formulate it as an MDP problem. However, it is difficult to solve this problem by using the traditional value iteration or policy iteration methods [28]. First, its state space is uncountable or approximately continuous due to the correlation among arms; second, some approximate solutions still suffer from high computational complexity and lack of insights [16]–[19]. To overcome these, we further model this problem as a correlated restless multi-armed bandit (CRMAB) framework [29]. In this CRMAB problem, the player is the BS, and arms are the correlated IoT devices; while rewards and states are both the AoIs.

There are two main challenges in solving this CRMAB problem. First, the states of different sources are highly relevant. As a result, it is difficult to decouple the CRMAB problem into several single-armed bandit problems, which is a core step for the RMAB problem to reduce the computational complexity. Second, the system state space becomes uncountable or approximately continuous. The continuous state space may lack the structure feature, such as the semi-universal structure [30] and the monotonic structure [16], which is a critical information for the RMAB problem to establish the indexability and to drive the closed-form expression of the WI. To conquer these challenges, we introduce a pending state for each arm to approximately represent all the possible states that are resulted from other arms' actions. This pending state essentially captures the arms' correlations and the channels' conditions. In this way, we can successfully decouple the CRMAB into several independent single-armed bandit problems. After that, we establish the indexability of each single-armed bandit problem and derive its closed-form expressions of the generalized WI (GWI) for identical channel and the generalized partial WI (GPWI) non-identical channel. At last, a threshold-based scheduling policy can be constructed to solve this AoI-based minimization problem.

The difference between this work and the existing ones and the main contributions of this work are summarized as follows.

- We study the AoI-based minimization problem with the setting of stochastically identical and non-identical multi-channel, as well as the correlated sources. However, existing works are limited to either the case of independent sources or the identical multi-channel.
- We put forth the GWI-based and GPWI-based scheduling policies to solve the AoI-based minimization problem under the identical and non-identical channel, respectively.
- We provide two lower numerical performance bounds for the proposed policies by solving the relaxed Lagrange problem of the decoupled CRMAB problem.
- We conduct several simulations to evaluate the proposed polices. Numerical results validate these lower performance bounds and show that the proposed policies exhibit the best performance among the compared policies. More importantly, the adopted PCI model is more suitable for high-density networks than the setting of independent sources.

A. Related Work

Correlated sources in the context of AoI and IoT are investigated in [21]–[27], which can be roughly classified into two groups, i.e., the fully correlated information (FCI) model and the partially correlated information (PCI) model. The FCI model defines the case that multi-IoT devices' status information are required at the BS to finish a status update process. Specifically, an IoT device's status is successfully updated if and only if all the IoT devices that are correlated to this one have successfully updated their status at the same time. Ref. [21] studies the AoI minimization problem under the FCI model by modeling this problem as an episodic MDP and develops a deep reinforcement learning method to solve this problem. Similarly, ref. [22] investigates this problem in the wireless camera networks where images from different cameras are correlated with the overlapping fields of view (FoV). Thus, these correlated cameras are required to jointly update to the BS. In addition, ref. [23] considers an IoT monitoring system where only partial status information of the physical process can be observed by each IoT device. Therefore, the BS requires different IoT devices' samples to re-construct such a physical process.

By contrast, the PCI model defines the case when updating an IoT device's status; others correlated to this one will also reveal partial information. Thus, an unscheduled IoT device can also update a partial status if one of its correlated sources has successfully updated its status. This PCI model is studied in [24]–[27]. Ref. [24] introduces a probability p_c to model this correlation, defining the probability that a packet in one device will also bring updated information about other devices. Then, it proposes a queue-based scheduling policy to reduce

the sum AoI. Ref. [25] considers two correlated sources where one source's status update will influence another. By modeling this correlation as the covariance between the samples of the two sources, the authors propose an optimal-time based scheduling policy. Compared with the above works and our paper, authors in [26], [27] distinguish between source and IoT device. They investigate the correlation among IoT devices, i.e., multiple IoT devices are correlated when they observe the same source. Thus, when updating a device's status, others that have the same observation will also be updated. To solve this AoI-based minimization problem, they propose an MDP-based scheduling policy. However, in this paper, we consider the AoI-based minimization problem in IoT networks under the PCI model by formulating it as a CRMAB framework.

MAB is a basic framework for the sequential decision-making problem [31], where a decision-maker (or player) must select an arm from a set of arms with unknown distribution at each time round. After that, the player will observe a reward from the environment. According to the rewarding process, MABs can be roughly classified into stochastic bandits, adversarial bandits, and Markovian bandits (e.g., the RMAB)). Traditional MABs assume that the arms are independent. MABs with correlated arms are investigated in [32]–[37]. Refs. [32], [33] consider the combinatorial MAB problem where the correlated arms are aggregated into a super arm. The well-known unimodality MAB problem is considered in [34]–[36], where the mean rewards of the arms are assumed to have the unimodal structure or the quasi-concavity property. Ref. [37] investigates the MAB problem with a graphical structure in which the graph characterizes the correlation among arms. By taking advantage of this correlation, the above works can significantly reduce the total exploration time to accelerate the algorithm's convergence rate. However, these works are considered under the framework of stochastic bandits. By contrast, we study this problem under the RMAB framework.

RMAB problem is a subset of the Markovian bandits where states at each arm are considered and evolve with time whenever the arm is active or not [29]. There are three main concepts in the RMAB problem, i.e., decoupling, indexability, and WI. The decoupling operation is the core step in reducing computational complexity. While the indexability of an RMAB problem is often difficult to establish, especially when the structure information of the arm's states is unknown [30]. The WI measures how rewarding to activate an arm by given a particular state. If an RMAB problem is proved to be indexable, the WI can be derived in closed-form. Refs. [16]–[19] apply the RMAB to the AoI-based minimization problem in IoT networks. By exploiting the monotonic structure of the AoI, ref. [16] derives the closed-form expression of the WI by using the recursive iteration method and proposes a WI-based scheduling policy to solve the AoI minimization problem. Similar to the WI's derivation process in [16], ref. [17] puts forth a WI-based decentralized scheduling policy. The analysis of the asymptotically optimal scheduling policy is investigated in [18], [19] by using the fluid analysis. Ref. [19] extends the concept of indexability to the non-identical multi-channel and proposes a sum weighted index matching (SWIM) policy based on the concept of the partial index (PI).

However, there is no closed-form expression of the PI is given in [19]. In this paper, we apply the RMAB with correlated arms into the AoI-based minimization problem in IoT networks and derive the closed-form expression of the WI. The works most relevant to our work are [16] and [19]. However, they all assume that the sources are independent, ignoring the correlation among them. More importantly, ref. [16] derives the WI in the context of finite and integer states, while the states in our work are uncountable and non-integer. Note that the non-integer state brings a great challenge to derive the closed-form expression of the WI.

The remainder of this paper is organized as follows. In Section II, we introduce the system model and the MDP-based formulation. The definitions and the objective of the CRMAB framework are given in Section III. In Sections IV and V, we present the GWI-based and GPWI-based scheduling policies for the AoI minimization problem under the identical and non-identical channel models, respectively. Two lower numerical performance bounds of the proposed policies are given in Section VI. The numerical results are given in Section VII, and Section VIII concludes this paper.

II. System Model

We consider an IoT network with set $\mathcal{N}$ of N IoT devices[1] deployed in an area to monitor the environment, as shown in Fig. 1. These IoT devices need to report their samples, consisting of the devices' state status, to the BS through set $\mathcal{K}$ of K channels. We assume that these IoT devices are powered by batteries[2] and transmit with constant power. The time is slotted in $t = 1, 2, \cdots, T$. Assume that the number of IoT devices is larger than the available channels, and each channel can be occupied by at most one source at each time slot. At the beginning of each time slot, the BS decides which IoT devices should be scheduled to update their status through the K available channels. These sources' samples are correlated, i.e., when updating one's status, others correlated to this one will also reveal partial information. The system's goal is to minimize the average network-level AoI by scheduling K desirable IoT devices to update their status at each time slot.

A. Multi-Channel and Correlated Sources Model

We investigate two types of multi-channel models, i.e., the stochastically identical channel model and the stochastically non-identical channel model. Assume that each IoT device transmits to the BS with constant power P_{tr}. Then, the successful transmission probability that IoT device n transmits on channel k is defined as p_{nk}. The channel quality vector of IoT device n is denoted by $\vec{p}_n = (p_{n1}, \ldots, p_{nK})$. We assume that $p_{nk} \in (0, 1)$ is independent of source n and channel k.

[1]The IoT device is also referred to as source. Thus, IoT device and source are interchangeable in this paper.

[2]According to [38], IoT devices' batteries are carefully selected according to their surrounding environment, typically working for several months. Moreover, the battery life can be estimated in advance and replaced in time since the transmit power is fixed. Therefore, we do not consider the extreme case where an IoT device cannot update its status due to a low battery.

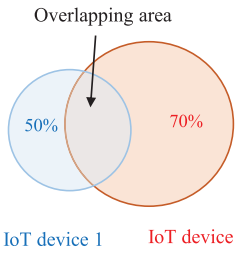


Fig. 2. A simple example illustrates the overlapped monitoring area among two IoT devices or cameras.

For the identical channel model, all p_{nk} are equal for $\forall k \in \mathcal{K}$, i.e., $p_{n1} = \ldots = p_{nK}$. As a result, here we can drop the index k when formulating the AoI-based minimization problem. However, in the non-identical channel model, the channels are potentially unreliable because the wireless channel suffers from diverse channel-fadings and geographical environment. Therefore, the successful transmission probabilities p_{nk} are different for $\forall k \in \mathcal{K}$ and $\forall n \in \mathcal{N}$.

According to [39], correlations among IoT devices can be characterized by their locations and samples, referring to the spatial correlation and the temporal correlation, respectively. In this paper, we define an undirected graph, $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, to capture the spatial and temporal correlations among the N IoT devices. In graph $\mathcal{G}$, the IoT device set $\mathcal{N}$ is the set of vertices, and $\mathcal{E}$ is the set of edges, which can be further written as an N-by-N matrix. Specifically, $\mathcal{E} = [\mathbf{e}_1; \mathbf{e}_2; \ldots; \mathbf{e}_N]$, where the k-th row vector is $\mathbf{e}_n = [e_{n1}, e_{n2}, \ldots, e_{nN}]$ and $\mathcal{E} \in [0,1]^{N \times N}$. Element e_{ij} is the j-th source's AoI-reduction factor, defined as the status update degree of source j when the BS receives the samples of source i. Thus, the diagonal elements are all equal to 1, i.e., $e_{ii} = 1, \forall i \in \mathcal{N}$. However, matrix $\mathcal{E}$ is asymmetric (i.e., $e_{ij} \neq e_{ji}$, $\forall i, j \in \mathcal{N}$) because the geographical location and the hardware accuracy of there IoT devices are non-identical. For example, in Fig. 2, the overlapping area accounts for 50% of IoT device 1's monitoring area and 30% of IoT device 2's monitoring area. Thus, the AoI-reduction factors of sources 1 and 2 are $e_{21} = 0.5$ and $e_{12} = 0.3$, respectively, which is asymmetric.

B. Age of Correlated Information Metric

Without loss of generality, we adopt the generate-at-will model as in [14], [16] to quantify the AoI evolution of the IoT network. Specifically, whenever an IoT device is scheduled, it immediately generates a packet for transmission. Let $A_n(t)$ be the age of IoT device n at time slot t. It reduces to 1 when IoT device n is scheduled and its transmission is successful; when it is not scheduled or its transmission is failed, its age is updated by the old value multiplied by an attenuation factor α_n and plus 1. Thus, the AoI evolution of IoT device n is

$$A_n(t+1) = \begin{cases} 1, & \text{TX success}, \\ A_n(t)\alpha_n(t) + 1, & \text{otherwise}, \end{cases} \tag{1}$$

where $\alpha_n(t) = \prod_{j \in \mathcal{I}_g^t}(1 - e_{jn})$ is the residual AoI degree of source n when the BS schedules the IoT device set $\mathcal{I}_g^t$ to update their status. Here, $\mathcal{I}_g^t$ denotes the set of IoT devices that are scheduled and their transmissions are successful at time slot t. For example, in Fig. 2, assuming that the AoIs of the IoT devices 1 and 2 are $A_1(t) = 10$ and $A_2(t) = 10$, respectively. When IoT device 1 is scheduled and its transmission is successful at time slot $t+1$, their AoIs are updated by $A_1(t+1) = 1$ and $A_2(t+1) = 10 \times (1 - 0.3) + 1 = 8$ according to (1). Note that, our AoI definition is different from that in [14], [16], [21]–[23], [27], as the status update of source n depends not only on itself but also on the scheduled sources that are related to source n.

Let $u_{nk}(t)$ be a binary decision variable at time slot t, where $u_{nk}(t) = 1$ means that source n is scheduled on channel k, and $u_{nk}(t) = 0$ otherwise. Thus, we have $\sum_{n=1}^{N} u_{nk}(t) = 1$, $\forall k \in \mathcal{K}$ since each channel can be occupied by at most one source at each time slot t. The system's goal is to minimize the average network-level AoI by finding the optimal scheduling policy π^* subject to the above constraints and correlation matrix $\mathcal{E}$ in graph $\mathcal{G}$, i.e.,

$$\min_{\pi} \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E}\left[A_n^{\pi}(t)\right] \tag{2a}$$

$$\text{s.t. } \alpha_n(t) = \prod_{j \in \mathcal{I}_g^t} (1 - e_{jn}), \quad \forall n \in \mathcal{N}, \ \forall t, \tag{2b}$$

$$\sum_{n=1}^{N} u_{nk}^{\pi}(t) = 1, \quad \forall k \in \mathcal{K}, \tag{2c}$$

$$\text{Eq. (1) and } u_{nk}^{\pi}(t) \in \{0,1\}, \quad \forall k \in \mathcal{K}, \ \forall t, \tag{2d}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator and π is the scheduling policy. The expectation is taken on the scheduling policy π. In the following, we refer to problem (2) as the original problem.

C. MDP-Based Formulation

It is not difficult to see that the original problem can be formulated as an MDP problem [10]. Let $\mathcal{A} \triangleq \{A_1(t), \ldots, A_N(t)\} \in \mathbb{N}_+^N$ be the system state space, i.e., the AoIs of N IoT devices at time slot t. Denote the action space of the entire system by $\mathcal{U} \triangleq \{u_{1k}, \ldots, u_{Nk}\} \in \{0,1\}^N$. According to (1), the state transition probability that the system state transits from $\mathbf{a}_t = \{A_1(t), \ldots, A_N(t)\}$ to $\mathbf{a}_{t+1} = \{A_1(t+1), \ldots, A_N(t+1)\}$ under action $\mathbf{u}_t \in \mathcal{U}$ is

$$\mathbb{P}_{\mathbf{a}_t|\mathbf{a}_{t+1}}(\mathbf{u}_t) = \begin{cases} \prod\limits_{n,k \in \mathcal{I}^t} p_{nk}, & \text{TX success}, \\ \prod\limits_{n,k \in \mathcal{I}^t} (1 - p_{nk}), & \text{TX failure}, \\ \prod\limits_{n,k \in \mathcal{I}_g^t} p_{nk} \prod\limits_{n,k \in \overline{\mathcal{I}}_g^t} (1 - p_{nk}), & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathcal{I}^t$ is the set of (n, k) pairs that source n is scheduled in channel k, i.e., $\mathcal{I}^t = \{(n,k) | u_{nk} = 1\}$. In addition, $\mathcal{I}_g^t$ represents the set of (n,k) pairs that source n is scheduled on channel k and its transmission is successful; while $\overline{\mathcal{I}}_g^t$ denotes the set of (n,k) pairs that source n is scheduled on channel k but its transmission is failed. Hence, we have $\mathcal{I}^t = \mathcal{I}_g^t \bigcup \overline{\mathcal{I}}_g^t$.

The objective of the MDP problem is to minimize the long-term average network-level AoI over time horizon T, i.e.,

$$\min_{\pi\in\mathcal{U}} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n^{\pi}(t)\right] \quad \text{s.t. (2b) (2c) and (2d).} \tag{4}$$

It is difficult to solve this problem by using the traditional value iteration or policy iteration method since the number of system states is uncountable or continuous. There are some works that discuss the continuous-state space MDP problem by discretizing the continuous state space, such as the relative value iteration (RVI) method [40]. Meanwhile, the other works tackle this problem by using the function approximation for the action value, such as the linear function approximation [41] and the neural network function approximation [42]. However, these solutions suffer from some performance loss or high computational complexity. In the following, we introduce the CRMAB framework to handle this problem.

III. Correlated RMAB Formulation

We first model the original problem as a CRMAB problem. Then, we give the concept of negative subsidy for passive and the objective of the CRMAB problem.

A. CRMAB-Based Formulation

As mentioned before, the RMAB problem can reduce the computational complexity by decoupling an N-dimensional problem into N independent 1-D problems when computing the WI. Then, a closed-form expression of the WI exists if all N 1-D problems are indexable. Therefore, one needs to prove the indexability of an RMAB problem first before solving it. For an indexable RMAB problem, a WI-based policy can be constructed by activating those K arms with the largest WIs at each time slot. If the WI calculation of the 1-D problem only relies on itself, we refer to it as a strongly decomposable WI. Here, we investigate the weakly decomposable WI to capture the correlations among sources. We will show that the strongly decomposable WI is a special case of the weakly decomposable WI. Therefore, we refer to the latter as generalized WI (GWI).

In the following, we model the original problem (2) as a CRMAB problem where the player is the BS, and the arms are the correlated sources. At the beginning of each time slot, the player decides which arms should be scheduled to update their state status. Then, the player will observe rewards $\vec{A}(t) = \{A_1(t),\ldots,A_N(t)\}$ at the end of each time slot, i.e., the AoI evolution in (1). For convenience, let $d_{n,t}$ denote the state of arm n at time slot t. Note that the state at each arm is the same as the reward, i.e., $A_n(t) = d_{n,t}$. According to the selection action $u_{nk}(t)$, state $d_{n,t}$ can be classified into the passive set $u_{nk}(t) = 0, \forall k \in \mathcal{K}$ and the active set $\sum_k u_{nk}(t) \neq 0$.

The state transition probability when source n is scheduled to channel k at time slot t and its transmission is failed is

$$\mathbb{P}\{A_n(t+1) = d_{n,t}\alpha_n(t)+1 | A_n(t)=d_{n,t}, u_{nk}(t)=1\} = 1-p_{nk}, \tag{5}$$

while its transmission is successful is

$$\mathbb{P}\{A_n(t+1)=1 | A_n(t)=d_{n,t}, u_{nk}(t)=1\} = p_{nk}. \tag{6}$$

In addition, the state transition probability when source n is not scheduled is given by

$$\mathbb{P}\{A_n(t+1)=d_{n,t}\alpha_n(t)+1 | A_n(t)=d_{n,t}, u_{nk}(t)=0\}=1, \tag{7}$$

where $\{A_n(t+1)|A_n(t), u_{nk}(t)\}$ denotes the state transition process of source n that state $A_n(t)$ moves to state $A_n(t+1)$ by taking the action $u_{nk}(t)$.

However, it can be seen from (5)-(7) that the state at each arm still depends on the others' actions (i.e., α_n), resulting in uncountable or continuous state space. To overcome this, we introduce an average residual AoI degree $\bar{\alpha}_n \in (0,1]$ for each arm, which is given by

$$\bar{\alpha}_n = \mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right], \quad \forall n \in \mathcal{N}, \tag{8}$$

where the expectation is taken on the scheduling strategy over the time horizon T (i.e., set $\mathcal{I}_g^t$). Therefore, we can replace $\alpha_n(t)$ with $\bar{\alpha}_n$ in (5) and (7). As a result, there are only two possible states for each arm at next time slot $t+1$, i.e., 1 or $d_{n,t}\bar{\alpha}_n + 1$.

Next, we show why the residual AoI degree $\alpha_n(t)$ can be replaced by the average residual AoI degree $\bar{\alpha}_n$. First, the BS is interested in the trends of the AoI growth of different sources over time slot t, rather than the specific value of the AoI at each source [16]. This growth trend essentially captures the correlation among arms and the condition of different channels. Second, the IoT network will converge to a stationary state (i.e., the average network-level AoI is fixed) for a given scheduling policy π. After achieving this stationary state, the scheduling strategy $\mathcal{I}_g^t$ changes periodically, as well as the residual AoI degree $\alpha_n(t)$. Thus, the average residual AoI degree $\bar{\alpha}_n$ will be a constant and independent of the time slot t, i.e., $\mathbb{E}[\alpha_n(t)] = \bar{\alpha}_n$, revealing the AoI growth trends among sources. In the following, we refer to this problem that using $\bar{\alpha}_n$ in (5)-(7) as the decoupled CRMAB problem.

B. Objective of the Decoupled CRMAB Problem

We adopt the concept of *subsidy for passivity* as in [29], [30] to construct the objective of the decoupled CRMAB problem. Here, we use the *negative subsidy*[3] (i.e., $-m$) for passivity because our objective is to minimize the average network-level AoI. The infimum m that is required to move a state from the active set to the passive set measures the attractiveness of an arm to be scheduled. Therefore, the states at each arm can be divided into the active set and the passive set with a particular m. This reveals the role of the solution for solving an RMAB problem, i.e., activates K arms with the largest m at each time slot t.

[3]The negative subsidy is also known as tax [29].

Therefore, the objective of the decoupled CRMAB problem is to minimize the average network-level AoI plus the negative subsidy for passivity $-m$, i.e.,

$$\begin{aligned} &\min \lim_{T\to\infty} \sum_{n=1}^{N} \mathbb{E}\left[J_n\right] \\ &\text{s.t. } J_n = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(A_n\left(t\right) - m_k\left(1-u_{nk}(t)\right)\right) \\ &\quad \text{(2c) (2d) and (8)}, \end{aligned} \tag{9}$$

where m_k is the subsidy for passivity of channel k. After introducing the average residual AoI degree $\bar{\alpha}_n$, the above problem can be decoupled into N single-armed bandit sub-problems. In the following, we focus on each sub-problem and drop the index of source n.

IV. GWI-Based Scheduling Policy Under Stochastically Identical Channel

In this section, we first give the concepts of indexability and GWI for the CRMAB framework. Then, we prove that the CRMAB problem is indexable and derive the closed-form expression of the GWI. Finally, we propose the GWI-based scheduling policy to solve the problem (9).

A. Definitions of Indexability and GWI

In the following, we can drop the index of channel k in the identical channel model. Therefore, there are only two possible actions (i.e., active $u = 1$ or passive $u = 0$) for each arm. For convenience, we assume that $d_{n,t}$ and d are interchangeable. According to [29], the single-armed bandit problem can also be viewed as an MDP. Let $V_m(d)$ be the differential value function at the initial state d with negative subsidy $-m$, representing the minimum expected total rewards that can be accrued from a single-armed bandit process.

The Bellman equation of the single-armed bandit is

$$V_m\left(d\right) + J^* = \min\left[V_m\left(d;u=0\right), V_m\left(d;u=1\right)\right], \tag{10}$$

where J^* is the optimal average reward and $V_m\left(d;u\right)$ is the expected total rewards obtained by taking action u under the optimal policy. According to state transition probabilities (5)-(7) and the objective function in (9), the above Bellman equation can be rewritten as in Eq. (11), shown at the bottom of the next page, where the initial condition is $V_m(1) = 0$. Thereafter, we have the following definitions.

Definition 1 (Passive Set): Let $\mathcal{P}(m)$ be the set of states d that the single-armed bandit is optimal to passive with the subsidy m. It satisfies the following condition,

$$\mathcal{P}(m) = \left\{d : V_m\left(d;u=0\right) \le V_m\left(d;u=1\right)\right\}. \tag{12}$$

Definition 2 (Indexability): An arm is indexable if the passive set $\mathcal{P}(m)$ of the single-armed bandit process monotonically increases from $\emptyset$ to the whole state space as m increases from $-\infty$ to $+\infty$. A CRMAB problem is indexable if and only if every arm is indexable.

Definition 3 (GWI): If an arm is indexable, its GWI $W(d,\bar{\alpha})$ is the infimum subsidy m that makes the scheduling decisions (active, passive) equally desirable at state d, i.e.,

$$W\left(d,\bar{\alpha}\right) = \inf_{m}\left\{m : V_m\left(d;u=0\right) = V_m\left(d;u=1\right)\right\}. \tag{13}$$

B. GWI-Based Scheduling Policy

We next investigate the indexability of the CRMAB problem in (9) and derive the closed-form expression of the GWI by solving the Bellman equation (10) or (11). As pointed out in [16], [30], the optimal solution for this Bellman equation is a threshold policy, i.e.,

Proposition 1: For the single-armed bandit problem with the average residual AoI degree $\bar{\alpha} \in (0,1]$, the optimal scheduling policy for solving the Bellman equation in (10) or (11) is a threshold policy. An arm is scheduled when $d \ge D$ and passive when $1 \le d < D$, where the threshold D is given by

$$D = \begin{cases} \dfrac{\bar{\alpha}(1-p)(p-m)+(m-1)}{p\bar{\alpha}(2-p-(1-p)\bar{\alpha})}, & 0<\bar{\alpha}<1, \\ -\left(1-m+\dfrac{(1-p)^2}{p}\right), & \bar{\alpha}=1. \end{cases} \tag{14}$$

Meanwhile, the optimal average reward J^* is given by

$$J^* = \begin{cases} \dfrac{1+Z-pm\tilde{n}}{p\tilde{n}+1} + \dfrac{p(\tilde{n}+p+p\bar{\alpha}-D\bar{\alpha})}{(1+p\tilde{n})(1-\bar{\alpha})}, \\ \dfrac{\left(1-p^2\right)^2-(m^2+2)p^2+2mp}{2p(1-p)}, \bar{\alpha}=1, \end{cases} \tag{15}$$

where the upper part of (15) is obtained under the case of $0 < \bar{\alpha} < 1$ and

$$\begin{cases} \tilde{n} = \log_{\bar{\alpha}}\left(1-D(1-\bar{\alpha})\right) - 2, \\ Z = \dfrac{\bar{\alpha}(1-p)(Dp+1-p)}{1-(1-p)\bar{\alpha}}. \end{cases} \tag{16}$$

Proof 1: See Appendix A.

Remark 1: From *Proposition* 1, we can see that the threshold D is a linear function of subsidy m in the case $0 < \bar{\alpha} < 1$. In other words, threshold D increases when m changes from $-\infty$ to $+\infty$ as $0 < \bar{\alpha} \le 1$. Thus, there exists an m^* such that the passive set $\mathcal{P}(m) = \emptyset$, i.e., $D(m^*) = 1$. Therefore, it is sufficient to show that the passive set $\mathcal{P}(m)$ monotonically increases from $\emptyset$ to the whole state space as m increases from $-\infty$ to $+\infty$. According to *Definition* 2, we conclude that the CRMAB problem in (9) is indexable since every single-armed bandit n is indexable.

Based on this indexability, we can derive the closed-form expression of GWI $W\left(d,\bar{\alpha}\right)$ by isolating m from (14). According to *Definition* 3, $W(d,\bar{\alpha})$ is the infimum subsidy m that makes the scheduling decisions (active, passive) equally desirable in state d. Hence, threshold D must be $d\bar{\alpha}+1$. Substituting $D = d\bar{\alpha}+1$ to (14) yields

$$m(d,\bar{\alpha}) = \begin{cases} \left(p\bar{\alpha}(d+1) + \dfrac{1+pd\bar{\alpha}(1-p)}{1-\bar{\alpha}(1-p)}\right), & 0<\bar{\alpha}<1, \\ \dfrac{p^2+dp+1}{p}, & \bar{\alpha}=1. \end{cases} \tag{17}$$

Algorithm 1 GWI-Based Scheduling Policy Under the Stochastically Identical Channel Model

1: **Initialize:** the parameters of $K, N, T, p_n, \mathcal{E}, A_n(1) = 1$.
2: **for** each time slot $t = 1, 2, \ldots, T$ **do**
3: Compute each arm's GWI by using (17)
4: Schedule arms with K largest values of $W(d, \bar{\alpha})$
5: Update each arm's AoI according to (1)
6: Calculate $\bar{\alpha}$ according to (8) or (19)
7: **end for**

Let $W(d, \bar{\alpha}) = m(d, \bar{\alpha})$ be the GWI. In the following, we provide two methods to compute the average residual AoI degree $\bar{\alpha}$ for the centralized and decentralized systems.

First, when the scheduling actions and the correlation matrix $\mathcal{E}$ are known to the BS in the centralized system, the average residual AoI degree $\bar{\alpha}_n$ of bandit n can be calculated by using (8). We refer to this method as the *realtime* residual AoI degree.

Second, the average residual AoI degree $\bar{\alpha}_n$ can be estimated from the historical AoI $A_n(t)$ when the prior information is unknown in the distributed system. According to (1), the AoI evolution can be generalized as

$$A_n(t+1) = A_n(t)\bar{\alpha}_n + 1. \tag{18}$$

This is a linear equation with slope $\bar{\alpha}_n$ which needs to be estimated. Let $H_y = A_n(t+1) - 1$ for $t = \{1, 2, \ldots, t\}$ and $H_x = A_n(t) - 1$ for $t = \{1, 2, \ldots, t\}$, the estimated slope $\hat{\alpha}_n$ can be obtained by using the least squares estimation (LSE) method, i.e.,

$$\hat{\alpha}_n = \frac{\sum_{i=1}^{t} \left(H_{x_i} - \bar{H}_x\right)\left(H_{y_i} - \bar{H}_y\right)}{\sum_{i=1}^{t} \left(H_{x_i} - \bar{H}_x\right)^2}, \tag{19}$$

where $\bar{H}_x = \sum_{i=1}^{t} \left(A_n(i) - 1\right)/t$ and $\bar{H}_y = \sum_{i=2}^{t+1} \left(A_n(i) - 1\right)/t$. Notice that the future reward $A_n(t+1)$ can be predicted from its historical rewards $A_n(t)$ by using the linear prediction filter coefficients (LPC) method [43]. Thus, we refer to (19) as the *estimated* residual AoI degree.

Finally, we present the GWI-based scheduling policy for the CRMAB problem of (9), as shown in Algorithm 1. For any $0 < \bar{\alpha} \leq 1$, at each time slot t, the BS activates the arms with K largest values of the GWIs. Specifically, at the beginning of each time slot t, the BS calculates each arm's GWI using (17). Then, it activates the arms with K largest values of GWIs to update their state status. The BS updates each IoT device's AoI according to their transmission outcomes. At last, the average residual AoI degree $\bar{\alpha}_n$ is calculated by using (8) or (19). Algorithm 1 repeats these steps until it reaches the stopping time T.

V. GPWI-Based Scheduling Policy Under Stochastically Non-Identical Channel

We first give the definitions of the partial indexability and the GPWI in Section V-A. Then, we prove this problem is partial indexable and derive the closed-form expression of the GPWI. Finally, we propose a GPWI-based scheduling policy to solve the CRMAB problem (9) in Section V-B.

A. Definitions of Partial Indexability and GPWI

We start with the single-armed bandit process in the CRMAB problem of (9), which can be viewed as an MDP problem. Consequently, we can drop the index of source n. In the non-identical channel model, each IoT device has $K+1$ possible actions to choose from. Specifically, $u = \{0, 1, 2, \ldots, K\}$, where $u = 0$ means that this bandit is passive. The Bellman equation of this single-armed bandit process can be written as

$$\begin{aligned} V_m(d) + J^* &= \min_{u \in \{0,1,\ldots,K\}} V_m(d, u) \\ &= \min_{u \in \{0,1,\ldots,K\}} \left\{ r_u(d) + \sum_{d'} \mathbb{P}_{d'd}(u) V_{\vec{m}}(d') \right\}, \end{aligned} \tag{20}$$

where $\mathbb{P}_{d'd}(u)$ is the state transition probability that the bandit transfers from state d to state d' when taking action u. Term $r_u(d)$ denotes the sum of the immediate reward and the negative subsidy $-m$ by taking action u under state d, and $\vec{m} = \{m_1, m_2, \ldots, m_K\}$ is the subsidy vector for the K channels.

Compared with the identical channel model, here, an IoT device that is scheduled in different channels will have distinct rewards at time slot t. According to (20), the decision of choosing channel k depends not only on the subsidy of this channel but also on the subsidies of others. Therefore, there is no longer a single threshold (or subsidy) that divides the state spaces into passive set and active set. Consequently, the Bellman equation can be rewritten as,

$$\begin{aligned} &V_{\vec{m}}(d) + J^* \\ &\quad = \min_{u \in \{0,1,\ldots,K\}} V_{\vec{m}}(d, u) \\ &\quad = \min_{u \in \{1,\ldots,K\}} \min_{k} \left[V_{m_k}(d; u{=}0), V_{m_k}(d; u{=}k) \right], \end{aligned} \tag{21}$$

where $V_{m_k}(d; u = 0)$ is the sum of the subsidy m_k and the immediate reward plus the total future rewards when taking action $u = 0$ with state d on channel k.

In fact, Eq. (21) can be decomposed into K independent single-armed bandit processes with actions $u = 0$ and $u = 1$ since the K non-identical channels are stochastically independent. The proof is given in Appendix B. The single-armed bandit process can be regarded as an MDP. As a result, the Bellman equation of the above MDP problems can be

$$V_m(d) + J^* = \min \begin{cases} V_m(d; u = 0) = d\bar{\alpha} + 1 - m + V_m(d\bar{\alpha} + 1), \\ V_m(d; u = 1) = p + (d\bar{\alpha} + 1)(1 - p) + V_m(1)p + V_m(d\bar{\alpha} + 1)(1 - p), \end{cases} \tag{11}$$

solved independently as the same in the identical channel model. Before presenting the main results in this section, we give the following definitions by generalizing the concepts in Section IV to the non-identical channel model [19].

Definition 4 (Passive Set): Given the subsidy vector $\vec{m}$, the passive set $\mathcal{P}_k(\vec{m})$ is the set of states d such that it is optimal to passive for channel k. It satisfies the following equation,

$$\mathcal{P}_k(\vec{m}) = \{d : V_{m_k}(d) > \min_{u \neq k} V_{m_u}(d)\}, \tag{22}$$

where $u \in \{0, 1, 2, \ldots, K\}$.

Let $\vec{m}_{-k}$ be the subsidy vector of all channels except for channel k. Meanwhile, let $\vec{m}' = [m'_k, \vec{m}_{-k}]$ be a new subsidy vector by fixing all the subsidies in $\vec{m}_{-k}$, but changing the value of m'_k. Then, the partial indexability is defined by

Definition 5 (Partial Indexability): Given the subsidy vector $\vec{m}$ and fixing the subsidy vector $\vec{m}_{-k}$, an arm is partially indexable for channel k if the passive set $\mathcal{P}_k(\vec{m}')$ of the single-armed bandit process increases from $\emptyset$ to the whole state space as m'_k increases from $-\infty$ to $+\infty$. A CRMAB problem is partially indexable if and only if all arms in the K non-identical channels are partially indexable.

Definition 6 (GPWI): Given the subsidy vector $\vec{m}$ and fixing the subsidy vector $\vec{m}_{-k}$, if an arm is partially indexable for channel k, its GPWI $G_k(d, \bar{\alpha})$ is the infimum subsidy m_k that makes the scheduling decisions (active, passive) equally desirable at state d, i.e.,

$$G_k(d, \bar{\alpha}) = \inf_{m'_k} \left\{ m'_k : V_{m'_k}(d; u = 0) = V_{m'_k}(d; u = k) \right\}. \tag{23}$$

Unlike GWI, the GPWI of channel k is defined by all channels' subsidies rather than only channel k. Hence, the GPWI is an $N \times K$ matrix under the non-identical channel model for the decoupled CRMAB problem.

B. GPWI-Based Scheduling Policy

In the non-identical channel model, the optimal solution for the Bellman equation in (21) under channel k is also a threshold policy. Hence, we have the following proposition.

Proposition 2: Given the subsidy vector $\vec{m}$ and fixing the subsidy vector $\vec{m}_{-k}$, for the single-armed bandit problem, the optimal scheduling policy for solving the Bellman equation (21) under channel k is a threshold policy. An arm is scheduled when $d \geq D_k$ and passive when $1 \leq d < D_k$, where the threshold D_k is given by

$$D_k = \begin{cases} \dfrac{\bar{\alpha}(1-p_k)(p_k - m_k) + (m_k - 1)}{p_k \bar{\alpha}(2 - p_k - (1-p_k)\bar{\alpha})}, & 0 < \bar{\alpha} < 1, \\ -\left(1 - m_k + \dfrac{(1-p_k)^2}{p_k}\right), & \bar{\alpha} = 1. \end{cases} \tag{24}$$

Meanwhile, the optimal average reward J_k^* is given by

$$J_k^* = \begin{cases} \dfrac{1 + Z_k - p_k m_k \tilde{n}_k}{p_k \tilde{n}_k + 1} + \dfrac{p_k(\tilde{n}_n + p_k + p_k\bar{\alpha} - D_k\bar{\alpha})}{(1 + p_k\tilde{n})(1-\bar{\alpha})}, \\ \dfrac{\left(1 - p_k^2\right)^2 - (m_k^2 + 2)p_k^2 + 2m_k p_k}{2p_k(1-p_k)}, \bar{\alpha} = 1, \end{cases} \tag{25}$$

where the upper part of (25) is obtained under the case of $0 < \bar{\alpha} < 1$ and

$$\begin{cases} \tilde{n}_k = \log_{\bar{\alpha}}(1 - D_k(1-\bar{\alpha})) - 2, \\ Z_k = \dfrac{\bar{\alpha}(1-p_k)(D_k p_k + 1 - p_k)}{1 - (1-p_k)\bar{\alpha}}. \end{cases} \tag{26}$$

Proof 2: See Appendix B.

Remark 2: Following the same argument in Subsection IV-B, we can conclude that the CRMAB problem in (9) is partially indexable since every bandit n in different channel k is partially indexable.

Also, by substituting $D = d\bar{\alpha} + 1$ to (24), we have

$$m_k(d, \bar{\alpha}) = \begin{cases} \left(p_k\bar{\alpha}(d+1) + \dfrac{1 + p_k d\bar{\alpha}(1-p_k)}{1 - \bar{\alpha}(1-p_k)}\right), & 0 < \bar{\alpha} < 1, \\ \dfrac{p_k^2 + dp_k + 1}{p_k}, & \bar{\alpha} = 1. \end{cases} \tag{27}$$

According to *Definition 6*, the GPWI for bandit n under channel k is given by

$$G_{nk}(d, \bar{\alpha}) = \begin{cases} \left(p_{nk}\bar{\alpha}(d+1) + \dfrac{1 + p_{nk} d\bar{\alpha}(1-p_{nk})}{1 - \bar{\alpha}(1-p_{nk})}\right), \\ \qquad 0 < \bar{\alpha} < 1, \\ \dfrac{p_{nk}^2 + dp_{nk} + 1}{p_{nk}}, \quad \bar{\alpha} = 1, \end{cases} \tag{28}$$

where $\bar{\alpha}$ is the average residual AoI degree which can be calculated by (8) or (19).

We now return to the decoupled CRMAB problem in (9). Based on the above GPWIs, this problem can be transformed into a maximum weighted matching (MWM) problem, i.e.,

$$\begin{aligned} \max_{u_{nk} \in \{0,1\}} & \sum_{n=1}^{N} \sum_{k=1}^{K} G_{nk} u_{nk} \\ \text{s.t.} & \sum_{n=1}^{N} u_{nk} = 1, \quad \forall k \in \mathcal{K} \\ & \sum_{k=1}^{K} u_{nk} \leq 1, \quad \forall n \in \mathcal{N}. \end{aligned} \tag{29}$$

Therefore, the GPWI-based scheduling policy is performed by activating the arms with $u_{nk} = 1$ at each time slot t, as given in Algorithm 2.

VI. Performance Analysis

In this section, we give two lower numerical performance bounds for the proposed policies under the stochastically identical and non-identical channel models. These lower performance bounds are obtained by solving the Lagrangian problem of the relaxation version of the decoupled CRMAB problem in (9). In the CRMAB problem, the number of IoT devices that should be scheduled at each time slot is strictly limited to K. However, in the relaxed problem, we allow that the number of long-term (i.e., the total time horizon T) average scheduled

Algorithm 2 GPWI-Based Scheduling Policy Under the Stochastically Non-Identical Channel Model

1: **Initialize:** the parameters of $K, N, T, p_n, \mathcal{E}, A_n(1) = 1$.
2: **for** each time slot $t = 1, 2, \ldots, T$ **do**
3: Compute each arm's GPWI by using (28)
4: Obtain the decision variable u_{nk} by solving the MWM problem in (29)
5: Schedule arms according to the decision variable u_{nk}
6: update each IoT device's AoI according to (1)
7: Calculate $\bar{\alpha}$ according to (8) or (19)
8: **end for**

IoT devices is equal to K. This indicates that the achievable performance by solving the relaxed problem will be better than that of the GWI-based and GPWI-based policies, providing a lower performance bound for the decoupled CRMAB problem.

We first present the relaxed problem and its Lagrange dual problem. The original problem is rewritten as

$$\min_{d} \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\right] \tag{30a}$$

$$\text{s.t. } \bar{\alpha}_n = \mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right], \quad \forall n \in \mathcal{N}, \forall t, \tag{30b}$$

$$\sum_{n=1}^{N} u_{nk}(t) = 1, \quad \forall k \in \mathcal{K}, \tag{30c}$$

$$\text{Eq. (1) and } u_{nk}(t) \in \{0,1\}, \quad \forall t. \tag{30d}$$

To be consistent with the concept of negative subsidy for passive, the constraint (30c) can be transformed into $\sum_{n=1}^{N}(1-u_{nk}(t)) = N-1$. Hence, the relaxed problem is given by

$$\begin{aligned}&\min_{d} \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\right]\\ &\text{s.t. } \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[(1-u_{nk}(t))\right] = N-1, \quad \forall k \in \mathcal{K},\\ &\quad \text{(30b) and (30d).}\end{aligned} \tag{31}$$

Then, the Lagrangian function of (31) is defined as

$$\begin{aligned}L(d,\vec{m}) = \inf_{d}\sum_{n=1}^{N}\Bigg(&\frac{1}{T}\sum_{t=1}^{T}(\mathbb{E}\left[A_n(t)\right]\\ &-\sum_{k=1}^{K}m_k\left(1-u_{nk}(t)\right))\Bigg)+\sum_{k=1}^{K}m_k(N-1),\end{aligned} \tag{32}$$

where $\vec{m} = \{m_1, m_2, \ldots, m_K\}$ is the Lagrange multiplier vector for the K distinct channels. Note that the Lagrange multiplier m_k is equivalent to the subsidy m_k but has different physical meanings.

Mathematically, it is difficult to solve problem (32) as the infimum operates on the whole system state space. This space can be infinity when T and N are sufficiently large because of the correlation among arms. However, we observe that the first term of the objective function in (32) is essentially the decoupled CRMAB problem of (9). Hence, we can replace the first term with $\sum_{n=1}^{N} J_n^*$, where the expression of J_n^* is given in (15) for the identical channel model and in (25) for the non-identical channel model. As a result, we can obtain the Lagrange dual problem of (31). By solving this dual problem, we further provide a lower performance bound for the relaxed problem (31).

For the identical channel model, each IoT device only has two possible actions. Thus, the Lagrange dual problem of (31) can be written as

$$\begin{aligned}&\max_{m}\ \Gamma(m)\\ &\text{s.t. } \Gamma(m) = \sum_{n=1}^{N} J_n^* + m(N-K)\\ &\quad \text{(30b) and (30d).}\end{aligned} \tag{33}$$

It is easy to solve this dual problem since J_n^* in (15) is a concave function of m (*please see the proof in Appendix* C). Hence, the optimal solution m^* can be obtained by solving $\partial G(m)/\partial m = 0$, i.e.,

$$m^* = \begin{cases}\displaystyle\sum_{n=1}^{N}\frac{\partial J_n^*}{\partial m} + (N-K) = 0, & 0 < \bar{\alpha} < 1,\\ \dfrac{(N-K)\sum_{n=1}^{N}(1-p_n)+N}{\sum_{n=1}^{N}p_n}, & \bar{\alpha} = 1,\end{cases} \tag{34}$$

where m^* is a numerical value in the case of $0 < \bar{\alpha} < 1$ and can be obtained by using the Newton method.

For the non-identical channel model, there are total $K+1$ possible actions for each IoT device. According to Appendix B, the Bellman equation in (21) can be decomposed into K independent single-armed bandit processes with actions $u = 0$ and $u = 1$. As a result, the Lagrangian function in (32) can be transformed into

$$\begin{aligned}L(d,\vec{m}) &= \sum_{n=1}^{N}\sum_{k=1}^{K}J_{nk}^* + \sum_{k=1}^{K}m_k(N-1)\\ &= \sum_{k=1}^{K}\left(\sum_{n=1}^{N}J_{nk}^* + m_k(N-1)\right),\end{aligned} \tag{35}$$

where J_{nk}^* is given in (25). It is equivalent to solve the following Lagrange dual sub-problem of channel k since the K distinct channels are independent. Thus, we have

$$\begin{aligned}&\max_{m_k}\ \Gamma(m_k)\\ &\text{s.t. } \Gamma(m_k) = \sum_{n=1}^{N} J_{nk}^* + m_k(N-1)\\ &\quad \text{(30b) and (30d).}\end{aligned} \tag{36}$$

Note that Eq. (36) can be regarded as the AoI-based minimization problem with only one available channel. As a result, we can solve it using the same method as that in the identical

Algorithm 3 Computing the Lower Performance Bounds Under Both Channel Models

1: **Initialize:** the parameters of $K, N, T, p_n, \mathcal{E}, A_n(1) = 1$.
2: **for** each time slot $t = 1, 2, \ldots, T$ **do**
3: Compute each IoT's GWI or GPWI using (17) or (28)
4: Obtain the multiplier m^* or m_k^* using (34) or (37)
5: Schedule the IoT devices whose GWIs or GPWIs over m^* or $\max_k \vec{m}^*$
6: Update each arm's AoI according to (1)
7: Calculate $\bar{\alpha}$ according to (8) or (19)
8: **end for**

channel model. Therefore, the optimal Lagrange multiplier for channel k is

$$m_k^* = \begin{cases} \sum_{n=1}^{N} \frac{\partial J_{n,k}^*}{\partial m_k} + N - 1 = 0, & 0 < \bar{\alpha} < 1, \\ \frac{(N-1)\sum_{n=1}^{N}(1-p_{nk}) + N}{\sum_{n=1}^{N} p_{nk}}, & \bar{\alpha} = 1. \end{cases} \tag{37}$$

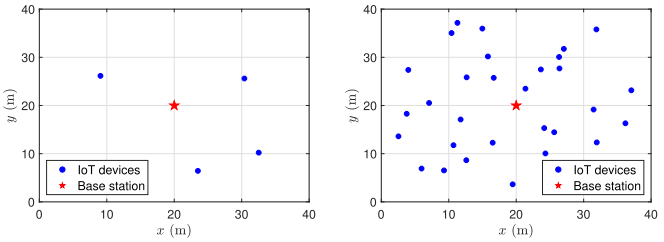
By solving the K sub-problems, we can obtain the optimal Lagrange multiplier vector $\vec{m}^* = \{m_1^*, m_2^*, \ldots, m_K^*\}$.

This relaxed problem reveals the role of subsidy m as the Lagrange multiplier and the asymptotically optimality of the proposed policies for the decoupled CRMAB problem. First, under the relaxed constraint, the proposed policies are implemented by activating those arms whose indexes at current states are over a constant m^* at each time slot. Second, the constant m^* is the Lagrange multiplier that makes the relaxed constraint satisfied, or that achieves the maximum in the dual problem in (33) and (36). Moreover, according to the optimization theory [44], the solution of the Lagrange dual problem (i.e., m^*) provides a lower performance bound to the relaxed problem. Therefore, Eqs. (34) and (37) are asymptotically optimal policies for the problem (9).

Finally, we give an algorithm to compute these lower performance bounds, as shown in Algorithm 3. At each time slot t, the BS calculates each IoT device's GWI or GPWI by using (17) or (28). Meanwhile, the optimal Lagrange multiplier m^* or $\vec{m}^*$ is calculated according to (34) or (37). Then, the BS compares the GWIs with m^* for the identical channel model, or compares the GPWIs with m_k^* for the non-identical channel model. It schedules the IoT devices whose indexes are over than m^* or $\max_k \vec{m}^*$. After transmissions, the BS updates each IoT device's AoI according to (1). Based on the updated AoIs and the correlation matrix $\mathcal{E}$, the average residual AoI degree $\bar{\alpha}$ can be calculated by using (8) or (19). Algorithm 3 repeats these steps until reaching the stopping time T.

VII. Simulation Results

We conduct several simulations to evaluate the performance of the GWI-based and GPWI-based scheduling policies with different network settings. The simulation parameters are chosen from the 3GPP standard [45]. All numerical results are obtained from 10^4 Monte Carlo (MC) trials.



(a) Network scenario I (b) Network scenario II

Fig. 3. The IoT network scenario in a (40×40) m^2 area with $N = 4$ IoT devices in (a) and $N = 30$ IoT devices in (b).

We first consider two IoT network scenarios in a (40×40) m^2 square area, as shown in Fig. 3. The number of IoT devices in network scenario I (Fig. 3a) and network scenario II (Fig. 3b) are $N = 4$ and $N = 30$, respectively. The locations of N IoT devices are generated by using the homogenous Poisson point process (HPPP) with user density parameter $\lambda = N/(\pi R^2)$, where $R = 20$ m is the radius of the largest circle in this square area.

We adopt the Indoor-Mixed office model [45] to capture the LoS component. Thus, the successful transmission probability of IoT device n is given by

$$p_n = \begin{cases} 1, & L_{n,n_0} \leq 1.2\text{m}, \\ \exp\left(-\frac{L_{n,n_0} - 1.2}{4.7}\right), & 1.2\text{m} < L_{n,n_0} \leq 6.5\text{m}, \\ \exp\left(-\frac{L_{n,n_0} - 6.5}{32.6}\right) \cdot 0.32, & 6.5\text{m} \leq L_{n,n_0}, \end{cases} \tag{38}$$

where L_{n,n_0} is the Euclidean distance between IoT device n and BS n_0. It is sufficient to model the identical channel model by only considering the LoS component. However, for the non-identical channel model, we consider the NLoS component in the successful transmission probability. We use the random variable ξ_{nk} to model NLoS component that IoT device n is scheduled on channel k, i.e., $p_{nk} = \xi_{nk} p_n$ where $\xi_{nk} \in (0, 1)$ follows the uniform distribution.

For the PCI model, the spatial correlation[4] among two sources is assumed to be inversely proportional to their distance [39]. Therefore, element e_{ij} in correlation matrix $\mathcal{E}$ can be calculated by $e_{ij} = \omega_i \exp(-\kappa L_{i,j})$, where $\omega_i \in (0, 1)$ is the weighting factor of source i, reflecting the importance of source i in the IoT network. Term $L_{i,j}$ is the Euclidean distance between sources i and j. In addition, κ is the correlation parameter that controls the strength of correlation, i.e., a bigger (or smaller) κ corresponds to a weaker (or higher) correlation. In the following, we set κ to 0.05. Notice that matrix $\mathcal{E}$ is asymmetric since ω_i is a random variable for different IoT devices.

Fig. 4 shows the average network-level AoI of the GWI-based and GPWI-based policies by using the realtime residual AoI degree $\bar{\alpha}$ and the estimated residual AoI degree $\hat{\alpha}$ in the network scenario I under the identical and non-identical

[4]Here we ignore the temporal correlations among sources to focus our discussion on the essence of the proposed scheduling policy.

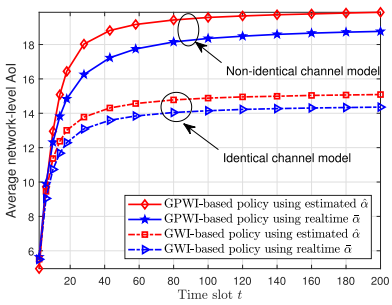
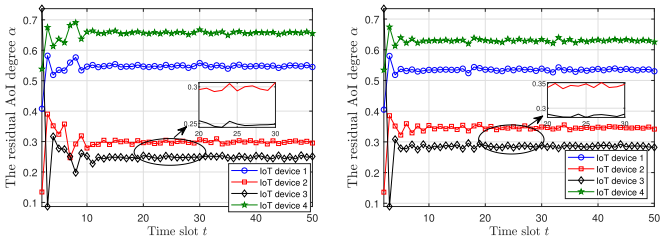


Fig. 4. The performance of the GWI-based policy with the realtime residual AoI degree $\bar{\alpha}$ and the estimated residual AoI degree $\hat{\alpha}$ in the network scenario I when $N = 4$ and $K = 2$.



(a) Using the estimated residual AoI degree $\hat{\alpha}$ (b) Using the realtime residual AoI degree $\bar{\alpha}$

Fig. 5. The residual AoI degree of different IoT devices v.s. time slot t in network scenario I when $N = 4$ and $K = 2$ by running the GWI-based policy with the estimated and realtime residual AoI degree.

channel models. The realtime $\bar{\alpha}$ is obtained by (8) and the estimated $\hat{\alpha}$ is obtained by (19). The number of available channels is $K = 2$ and the average network-level AoI is calculated by $\left(\sum_{i=1}^{t}\sum_{n=1}^{N} A_n(i)\right)/t$.

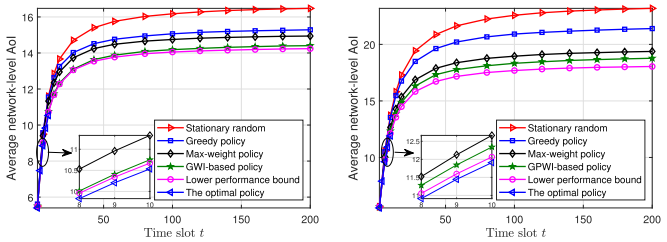
We can see from Fig. 4 that the performance of GWI-based and GPWI-based policies with the realtime $\bar{\alpha}$ is slightly better than that of estimated $\hat{\alpha}$ under both channel models. This is because that the realtime $\bar{\alpha}$ case has the prior information of the correlation matrix $\mathcal{E}$ and each time's scheduling actions. It can be also seen that the GWI-based policy outperforms the GPWI-based policy. The reason is that the non-identical channel is jointly modeled by the NLoS and LoS components; while the identical channel only considers the LoS component in the simulation setting. In other words, the successful transmission probability under the non-identical channel model is lower than that in the identical channel model. More importantly, Fig. 4 indicates that even though the coefficient matrix $\mathcal{E}$ is unknown prior, the GWI-based and GPWI-based policies can still work efficiently by using the estimated $\hat{\alpha}$. In order to compare, we evaluate the proposed policies by using realtime $\bar{\alpha}$ in the following.

Fig. 5 shows the residual AoI degree of different IoT devices v.s. time slot t in network scenario I when $N = 4$ and $K = 2$ by running the GWI-based policy using the estimated residual AoI degree $\hat{\alpha}$ and the realtime residual AoI degree $\bar{\alpha}$. We can see that the residual AoI degree $\alpha_n(t)$ of different IoT devices will trend to the stationary state (or changes periodically) in Figs. 5a and 5b. This demonstrates that $\mathbb{E}(\alpha_n(t)) = \bar{\alpha}_n$. Therefore, it is reasonable to replace $\alpha_n(t)$ with $\bar{\alpha}_n$ to decouple the CRMAB problem and to discretize the continuous state space.

TABLE I

INDEXES OF DIFFERENT SCHEDULING POLICIES ON IOT DEVICE n

Greedy	Random	Max-Weight	GWI	GPWI
$A_n(t)$	$\eta_n/\sum_{i=1}^{N}\eta_i$	$p_n A_n(t)(A_n(t)+2)$	Eq. (17)	Eq. (28)



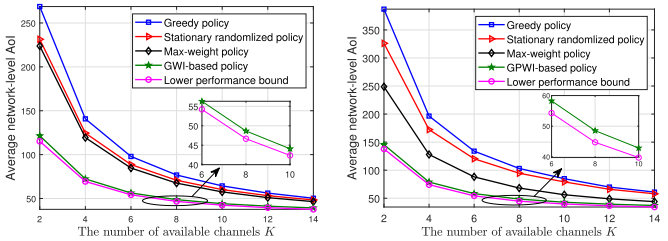
(a) Identical channel model (b) Non-identical channel model

Fig. 6. The average network-level AoI of different scheduling policies v.s. time slot t in network scenario I where $N = 4$ and $K = 2$.

Next, we compare different scheduling policies, such as the optimal policy, lower performance bound, greedy policy, stationary randomized policy, max-weight policy, GWI-based policy, and GPWI-based policy in terms of the average network-level AoI in network scenario I. The number of available channels is $K = 2$. The optimal policy is obtained by solving the MDP problem in (4) by using the value iteration method. However, since the number of states increases exponentially with time slot t and the number of IoT devices N, we can only compute the case of the time slot $t \leq 10$ and the number of IoT devices $N \leq 4$. In addition, the lower performance bound is obtained by running the Algorithm 3.

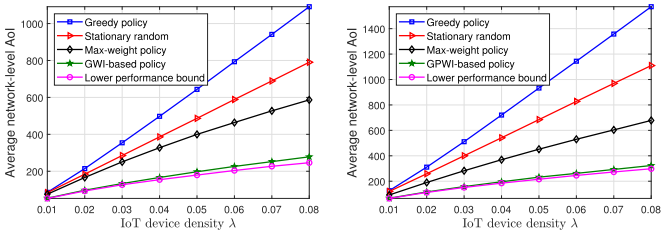
Other scheduling policies are summarized in Table I. The greedy policy schedules the IoT devices with K highest values of AoI $A_n(t)$ at each time slot t. The stationary randomized policy activates K IoT devices at each time slot t with the highest probabilities of $\eta_n/\sum_{i=1}^{N}\eta_i$, where η_n is a constant value associated with IoT device n. The max-weight policy is adopted according to [16], which is a function of AoI $A_n(t)$ and the transmission successful probability p_n. Specifically, it schedules the K IoT devices at each time slot t with the highest values of $p_n A_n(t)(A_n(t)+2)$. For the GWI-based and GPWI-based policies, we run Algorithms 1 and 2 to compute the average network-level AoI, respectively.

Fig. 6 depicts the performance of the above scheduling policies in network scenario I under the identical and non-identical channel model where $N = 4$ and $K = 2$. It can be seen that all scheduling policies can converge when the time slot t approaches $T = 200$ under both channel models. However, the performance of the proposed policies outperforms the existing scheduling policies, i.e., the greedy policy, the stationary randomized policy, the max-weight policy. Furthermore, the performance of the proposed policies are close to the lower performance bounds and the optimal solution. These results demonstrate that the GWI-based and GPWI-based policies not only consider the sources' own AoI and the successful transmission probability, but also take other IoT devices' actions into account (i.e., the parameter $\bar{\alpha}$).



(a) Identical channel model (b) Non-identical channel model

Fig. 7. The performance of different scheduling policies v.s. the number of available channels in network scenario II where $T = 500$ and $N = 30$.



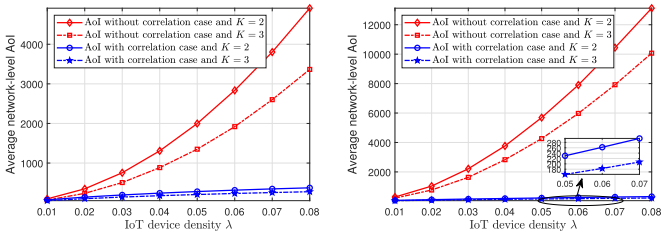
(a) Identical channel model (b) Non-identical channel model

Fig. 8. The performance of different scheduling policies v.s. the IoT device's density λ under the random network scenario where $T = 500$ and $K = 2$.

Under network scenario II, Fig. 7 presents the average network-level AoI of the above scheduling policies when the number of the available channels K is changed from 2 to 14, where the total time slot is $T = 500$ and the number of IoT devices is $N = 30$. It can be seen that the performance of all scheduling policies decreases with the number of available channels K under both channel models. Furthermore, the proposed policies have the best performance among the existing scheduling policies and very close to the lower performance bound under both two channel models. However, the performance gain is not significant when the number of available channels is over 8. Thus, there is a tradeoff between the performance gain and the cost of increasing the number of available channels.

In the following, we evaluate the GWI- and GPWI-based policies by considering different IoT device's density λ in random network scenario under both channel models. For the random network scenario, at each MC trial, the transmission successful probability $\vec{p}$ and the correlation matrix $\mathcal{E}$ are generated randomly as the same process of the generation of the network scenarios I or II. We also consider two possible network cases, i.e., the AoI with correlation case ($0 < \bar{\alpha} < 1$) and the AoI without correlation case ($\bar{\alpha} = 1$).

Fig. 8 compares the performance of different scheduling policies under the AoI with correlation case when IoT device's density λ changes from 0.01 to 0.08 under the random network scenario, where the total time slot is $T = 500$ and the number of available channels is $K = 2$. We can see that the proposed policies have the best performance among the existing scheduling policies and are close to the lower performance bound under both channel models. The performance gaps between the proposed policies and other policies increase with the IoT device density λ. In addition, the performance of the greedy policy accounts for the worst performance among all scheduling policies.



(a) Identical channel model (b) Non-identical channel model

Fig. 9. The performance of the GWI-based and GPWI-based policies v.s. the IoT device's density λ under the random network scenario in the cases of AoI with and without correlation where $T = 500$.

Fig. 9 shows the performance of the GWI-based and GPWI-based policies versus the IoT device's density λ under the AoI with correlation case and the AoI without correlation case in the random network scenario where the total time slot $T = 500$. The number of available channels is set to $K = 2$ and $K = 3$. It can be seen that, under both channel models, the average network-level AoI of the correlation case is much lower than that of the non-correlation case when $K = 2$ and $K = 3$. Moreover, the performance gap between these two cases increases with the IoT device density λ for $K = 2$ and $K = 3$. Fig. 9 demonstrates that the proposed policies can reduce the average network-level AoI significantly in the AoI correlation case under both channel models, which benefit the AoI-based scheduling problem in high-density networks.

VIII. Conclusion and Discussion

This paper studied the AoI-based minimization problem in IoT networks with the settings of identical and non-identical multi-channel, as well as correlated sources. We first formulated this problem as an MDP, but its solution suffers from high computational complexity and a lack of insights. Then, we modeled this problem as a CRMAB problem, which is computationally efficient and intuitively interpretable. We derived the closed-form expressions of the GWI and GPWI under the identical and non-identical channel models, respectively. As a result, the GWI-based and GPWI-based scheduling policies are constructed to solve this decoupled CRMAB problem. Moreover, we provided two lower numerical performance bounds for the proposed policies by solving the relaxed Lagrange problem. Numerical results validated the efficiency of the proposed scheduling policy and demonstrated that they outperform the state-of-the-art scheduling policies. More importantly, the adopted PCI model and the proposed policies can benefit the AoI-based scheduling problem in high-density networks.

In the CRMAB problem formulation, we have introduced a pending state for each arm, removing the dependency among sources to some extent. Then, an interesting problem is to find the optimal solution to the original problem (2) by solving the continuous-state space MDP problem in (4). This is an important yet challenging problem for future study.

APPENDIX A
PROOF OF PROPOSITION 1

We rewrite the Bellman equation of (11) as

$$V_m(d) = d\bar{\alpha} + 1 - J^* + V_m(d\bar{\alpha}+1) + \min\begin{cases} -m, \\ -pd\bar{\alpha} - pV_m(d\bar{\alpha}+1), \end{cases} \tag{39}$$

where the upper part of (39) is associated with the action of $u = 0$ and the bottom part is with $u = 1$. Let D be the threshold of the scheduling policy. Then, an arm is scheduled when $d \geq D$ and passive when $1 \leq d < D$.

Case I (We First Consider the Case of $0 < \bar{\alpha} < 1$): For $d \geq D$, we have $-pd\bar{\alpha} - pV_m(d\bar{\alpha}+1) < -m$, i.e.,

$$V_m(d\bar{\alpha}+1) > \frac{m - pd\bar{\alpha}}{p}. \tag{40}$$

Meanwhile, the value function of $V_m(d)$ is given by

$$V_m(d) = (1-p)V_m(d\bar{\alpha}+1) + (1-p)d\bar{\alpha} + 1 - J^*. \tag{41}$$

Then, we have the following recurrence relationships:

$$\begin{cases} V_m(d) = (1-p)V_m(d\bar{\alpha}+1) + (1-p)d\bar{\alpha} + 1 - J^*, \\ \quad (1-p)V_m(d\bar{\alpha}+1) = (1-p)^2 V_m(d\bar{\alpha}^2 + \bar{\alpha} + 1) \\ \quad + (1-p)^2(d\bar{\alpha}+1) + (1-p)(1-J^*), \\ \quad \vdots \\ (1-p)^n V_m(d\bar{\alpha}^n + \ldots + 1) = (1-p)^{n+1} V_m(d\bar{\alpha}^{n+1} + \ldots + 1) \\ \quad + (1-p)^{n+1}(d\bar{\alpha}^{n+1} + \ldots + \bar{\alpha} + 1) \\ \quad + (1-p)^n(1-J^*). \end{cases}$$

By summing over the above equations, we obtain

$$\begin{aligned} V_m(d) = {} & (1-p)^{n+1} V_m(d\bar{\alpha}^{n+1} + \ldots + 1) \\ & + (1-p)d\bar{\alpha} + \ldots + (1-p)^{n+1}(d\bar{\alpha}^{n+1} + \ldots + \bar{\alpha}) \\ & + (1-J^*)\left(1 + (1-p) + \ldots + (1-p)^n\right). \end{aligned}$$

Thus, when $n \to +\infty$, we have $(1-p)^{n+1} \to 0$ and

$$V_m(d) = \frac{\bar{\alpha}(1-p)(dp+1-p)}{p(1-(1-p)\bar{\alpha})} + \frac{1-J^*}{p}. \tag{42}$$

Since $1 \leq d < D$, we have $-pd\bar{\alpha} - pV_m(d\bar{\alpha}+1) \geq -m$, i.e.,

$$V_m(d\bar{\alpha}+1) \leq \frac{m - pd\bar{\alpha}}{p}. \tag{43}$$

Meanwhile, the differential value function of $V_m(d)$ is

$$V_m(d) = V_m(d\bar{\alpha}+1) - m + d\bar{\alpha} + 1 - J^*. \tag{44}$$

Substituting $d = (D-1)/\bar{\alpha}$ into (44) yields

$$V_m(D) = V_m\left(\frac{D}{\bar{\alpha}} - \frac{1}{\bar{\alpha}}\right) - D + (J^* + m). \tag{45}$$

Similarly, we have the following recurrence relationships:

$$\begin{cases} V_m(D) = V_m\left(\frac{D}{\bar{\alpha}} - \frac{1}{\bar{\alpha}}\right) - D + (J^* + m), \\ V_m\left(\frac{D}{\bar{\alpha}} - \frac{1}{\bar{\alpha}}\right) = V_m\left(\frac{D}{\bar{\alpha}^2} - \frac{1}{\bar{\alpha}^2} - \frac{1}{\bar{\alpha}}\right) \\ \quad - \left(\frac{D}{\bar{\alpha}} - \frac{1}{\bar{\alpha}}\right) + (J^* + m), \\ \quad \vdots \\ V_m\left(\frac{D}{\bar{\alpha}^n} - \frac{1}{\bar{\alpha}^n} - \ldots - \frac{1}{\bar{\alpha}}\right) = V_m\left(\frac{D}{\bar{\alpha}^{n+1}} - \frac{1}{\bar{\alpha}^{n+1}}\right. \\ \quad \left. - \ldots - \frac{1}{\bar{\alpha}}\right) - \left(\frac{D}{\bar{\alpha}^n} - \frac{1}{\bar{\alpha}^n} - \ldots - \frac{1}{\bar{\alpha}}\right) + (J^* + m). \end{cases}$$

By summing over the above equations, we obtain

$$\begin{aligned} & V_m\left(\frac{D}{\bar{\alpha}^{n+1}} - \frac{1-\bar{\alpha}^{n+1}}{\bar{\alpha}^{n+1} - \bar{\alpha}^{n+2}}\right) \\ & \quad = V_m(D) + n\left(\frac{1}{1-\bar{\alpha}} - J^* - m\right) + \frac{\bar{\alpha}^{n+1} - 1}{\bar{\alpha}^{n+1} - \bar{\alpha}^n} D \\ & \qquad + \frac{\bar{\alpha}^n - 1}{\bar{\alpha}^n (1-\bar{\alpha})^2}, \end{aligned} \tag{46}$$

where $V_m(D)$ is given in (42) and $1 \leq n < D$.

We have obtained the differential value functions as shown in (42) and (46) which are the function of threshold D, the optimal average reward J^*, and the subsidy for passive m. To find these variables, it is sufficient to investigate the conditions of $V_m(1) = 0$ and $V_m(D + J^*) = (m - pD\bar{\alpha})/p$. The latter condition comes from the fact that there is an optimal state $(D + J^*)$, where $J^* \in [0, 1)$, such that the inequalities of (40) and (43) hold. Thus, we obtain

$$V_m(D) \leq \frac{m - pD\bar{\alpha}}{p} < V_m(D\bar{\alpha}+1). \tag{47}$$

So there exists a $J^* \in [0, 1)$ such that $V_m(D + J^*) = (m - pD\bar{\alpha})/p$. According to the above two conditions, we have

$$\begin{cases} \frac{\bar{\alpha}(1-p)(1-p+(D+J^*p))}{1-(1-p)\bar{\alpha}} = J^* + m - pD\bar{\alpha} - 1, \\ \frac{-m + Dp\bar{\alpha}}{p} + \frac{(1-p)\bar{\alpha}J^*}{1-(1-p)\bar{\alpha}} = \tilde{n}\left(\frac{1}{1-\bar{\alpha}} - J^* - m\right) \\ \quad + \frac{\bar{\alpha}^{\tilde{n}+1} - 1}{\bar{\alpha}^{\tilde{n}+1} - \bar{\alpha}^{\tilde{n}}} D + \frac{\bar{\alpha}^{\tilde{n}} - 1}{\bar{\alpha}^{\tilde{n}}(1-\bar{\alpha})^2}, \end{cases} \tag{48}$$

where $\tilde{n}$ is obtained by solving $\frac{D}{\bar{\alpha}^{n+1}} - \frac{1-\bar{\alpha}^{n+1}}{\bar{\alpha}^{n+1} - \bar{\alpha}^{n+2}} = 1$. The solution of $\tilde{n}$ is given by

$$\tilde{n} = \log_{\bar{\alpha}}(1 - D(1-\bar{\alpha})) - 2, \tag{49}$$

where $\log_{\bar{\alpha}}(\cdot)$ is the logarithmic function with base $\bar{\alpha}$.

By combining (48), we obtain

$$D = \frac{\bar{\alpha}(1-p)(p - m - J^*p - J^*) + (J^* + m - 1)}{p\bar{\alpha}(2 - p - (1-p)\bar{\alpha})}. \tag{50}$$

Note that since

$$\frac{\partial D}{\partial J^*} = \frac{1 - (1-p^2)\bar{\alpha}}{p\bar{\alpha}(2 - p - (1-p)\bar{\alpha})} > 0, \tag{51}$$

$D(J^*)$ is monotonically increasing in the range $[0, 1)$. Hence, the optimal value can be achieved only in the end of the

interval, i.e., the expression D is obtained by letting $J^*=0$, which gives

$$D=\frac{\bar{\alpha}(1-p)(p-m)+(m-1)}{p\bar{\alpha}(2-p-(1-p)\bar{\alpha})}. \tag{52}$$

In addition, according to (48) and (52), we can obtain the optimal average reward

$$J^*=\frac{1+Z+pm\tilde{n}}{p\tilde{n}+1}+\frac{p(\tilde{n}+p+p\bar{\alpha}-D\bar{\alpha})}{(1+p\tilde{n})(1-\bar{\alpha})}, \tag{53}$$

where

$$Z=\frac{\bar{\alpha}(1-p)(DP+1-p)}{1-(1-p)\bar{\alpha}}. \tag{54}$$

Finally, we show that the solution of (52) is a threshold policy such that the following condition must be satisfied,

$$V_m\left(\frac{D}{\bar{\alpha}^{1+n^-}}-\frac{1-\bar{\alpha}^{1+n^-}}{\bar{\alpha}^{1+n^-}-\bar{\alpha}^{2+n^-}}\right)\le\frac{m-pD\bar{\alpha}}{p}<V_m(d'), \tag{55}$$

where $d'\in[D,+\infty)$ and $n^-\in\{0,1,2,\ldots\}$. Next, we show that the above condition holds. According to (42) and (46), it is easy to see that $V_m(d')$ and $V_m\left(D\bar{\alpha}^{n^-}+\frac{1-\bar{\alpha}^{n^-}}{1-\bar{\alpha}}\right)$ are monotonically increasing with d' and n^-, respectively. Therefore, the solution of (52) is a threshold policy that an arm is scheduled when $d\ge D$ and passive when $1\le d<D$.

Case II (We Next Investigate the Case of $\bar{\alpha}=1$): The main idea is the same as that in the case of $0<\bar{\alpha}<1$. The only difference lies in the geometric sequence summation operation of the recurrence relationships in (41) and (45). Hence, for the case of $\bar{\alpha}=1$, the differential value functions $V_m(d)$ can be recalculated by

$$\begin{cases} V_m(d)=\dfrac{p(1-J^*)+pd(1-p)+(1-p)^2}{p^2}, & d\ge D,\\ V_m(D-n)=V_m(D)+n\left(D-m-\dfrac{n}{2}\right), & 1\le d<D. \end{cases} \tag{56}$$

Similarly, by exploiting the conditions of $V_m(1)=0$ and $V_m(D+J^*)=(m-pD)/p$, we can compute the threshold expression D and the optimal reward J^*, respectively, i.e.,

$$\begin{cases} D=-\left(1-m+\dfrac{(1-p)^2}{p}\right),\\ J^*=\dfrac{\left(1-p^2\right)^2-(m^2+2)p^2+2mp}{2p(1-p)}. \end{cases} \tag{57}$$

At last, we need to verify that the solution of (57) is a threshold policy, which must satisfy the following condition,

$$V_m\left(D+n^-+1\right)\le\frac{m-pD}{p}<V_m(D+n^++1), \tag{58}$$

where $n^-\in\{-D+1,\ldots,-1\}$ and $n^+\in\{0,1,2,\ldots\}$. Next, we show that the above condition holds. According to (56), it is easy to see that $V_m\left(D+n^-\right)$ and $V_m\left(D+n^+\right)$ are monotonically increasing with n^- and n^+, respectively. Therefore, the solution of (57) is a threshold that an arm is scheduled when $d\ge D$ and passive when $1\le d<D$ in the case of $\bar{\alpha}=1$.

Appendix B
Proof of Proposition 2

Here, we just need to prove that the single-armed MDP problem can be decomposed into the multiple $\{0,1\}$-actions MDP problems since the threshold expression D_k is the same as in Proposition 1 except that it considers different channels. Then, each $\{0,1\}$-actions MDP problem can be solved efficiently as that in Proposition 1.

The Bellman equation of the single-armed MDP problem under non-identical channel can be rewritten as

$$V_m(d)+J^*=\min_{u\in\{0,1,\ldots,K\}}V_m(d,u), \tag{59}$$

where m is the subsidy for passive when the player takes action $u=0$. Since the K different channels are independent, the above equation can be transferred to

$$V_{\vec{m}}(d)+J^*=\min_u\begin{cases}\min\left[V_{m_1}(d;u=0),V_{m_1}(d;u=1)\right]\\ \min\left[V_{m_2}(d;u=0),V_{m_2}(d;u=2)\right]\\ \vdots\\ \min\left[V_{m_K}(d;u=0),V_{m_K}(d;u=K)\right],\end{cases}$$

where $\vec{m}=\{m_1,m_2,\ldots,m_K\}$ is the subsidy vector for the K distinct channels and $u\in\{1,\ldots,K\}$. Let $\vec{m}_{-k}$ be a subsidies vector of all channels except for channel k and let $\vec{m}'=[m'_k,\vec{m}_{-k}]$ be the new subsidy vector. We now fix all the subsidies in $\vec{m}_{-k}$, but change the value of m'_k. Then, the above equation is transferred to

$$\min_u\begin{cases}V_{\vec{m}'}(d)+J^*=\min\left[V_{m_1}(d;u=0),V_{m_1}(d;u=1)\right]\\ V_{\vec{m}'}(d)+J^*=\min\left[V_{m_2}(d;u=0),V_{m_2}(d;u=2)\right]\\ \vdots\\ V_{\vec{m}'}(d)+J^*=\min\left[V_{m_K}(d;u=0),V_{m_K}(d;u=K)\right].\end{cases}$$

As a result, Eq. (59) can be decomposed into K sub-problems of $\{0,1\}$-actions MDP. Therefore, it is not difficult to obtain the threshold expression of D_k by solving the $\{0,1\}$-actions MDP problem as that in Proposition 1. Finally, the proof is concluded by adopting the threshold expression of (14) and the optimal average reward expression of (15) in Proposition 1 for channel k.

Appendix C
Proof of the Concavity of J^* in (15)

We first prove the concavity of J^* under the case of $0<\bar{\alpha}<1$. It can be seen from (15) that the optimal average reward J^* is an affine function of the subsidy m. Thus, we can obtain that $J^*(m)$ is a concave function since its second-order derivative is equal to 0. For the case of $\bar{\alpha}=1$, we derive the second-order derivative of $J^*(m)$ which is given by

$$\frac{\partial^2J^*(m)}{\partial m^2}=\frac{-p}{1-p}. \tag{60}$$

Since $0<p<1$, we have $\partial^2J^*(m)/\partial m^2<0$. Hence, $J^*(m)$ is a strictly concave function in the case of $\bar{\alpha}=1$. Finally, we conclude that $J^*(m)$ is a concave function of m in (15) when $0<\bar{\alpha}\le1$.

REFERENCES

[1] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[2] H. Chen, Y. Gu, and S.-C. Liew, "Age-of-information dependent random access for massive IoT networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Jul. 2020, pp. 930–935.

[3] J. Tong, H. Zhang, L. Fu, A. Leshem, and Z. Han, "Two-stage resource allocation in reconfigurable intelligent surface assisted hybrid networks via multi-player bandits," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3526–3541, May 2022.

[4] M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "On the role of age of information in the Internet of Things," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 72–77, Dec. 2019.

[5] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Found. Trends Netw.*, vol. 12, no. 3, pp. 162–259, 2017.

[6] Z. Qian, F. Wu, J. Pan, K. Srinivasan, and N. B. Shroff, "Minimizing age of information in multi-channel time-sensitive information update systems," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Jul. 2020, pp. 446–455.

[7] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1770–1775.

[8] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Timely status update in massive IoT systems: Decentralized scheduling for wireless uplinks," 2018, *arXiv:1801.03975*.

[9] Q. He, G. Dan, and V. Fodor, "Joint assignment and scheduling for minimizing age of correlated information," *IEEE/ACM Trans. Netw.*, vol. 27, no. 5, pp. 1887–1900, Oct. 2019.

[10] B. Zhou and W. Saad, "Minimum age of information in the Internet of Things with non-uniform status packet sizes," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1933–1947, Mar. 2020.

[11] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "A reinforcement learning framework for optimizing age of information in RF-powered communication systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4747–4760, Aug. 2020.

[12] A. Javani, M. Zorgui, and Z. Wang, "Age of information in multiple sensing," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2020, pp. 1–10.

[13] V. Tripathi and E. Modiano, "Age debt: A general framework for minimizing age of information," 2021, *arXiv:2101.10225*.

[14] E. U. Atay, I. Kadota, and E. Modiano, "Aging bandits: Regret analysis and order-optimal learning algorithm for wireless networks with stochastic arrivals," 2020, *arXiv:2012.08682*.

[15] B. Sombabu and S. Moharir, "Age-of-information based scheduling for multi-channel systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4439–4448, Jul. 2020.

[16] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2637–2650, Dec. 2018.

[17] Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Can decentralized status update achieve universally near-optimal age-of-information in wireless multiaccess channels?" in *Proc. 30th Int. Teletraffic Congr. (ITC)*, Sep. 2018, pp. 144–152.

[18] A. Maatouk, S. Kriouile, M. Assad, and A. Ephremides, "On the optimality of the whittle's index policy for minimizing the age of information," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1263–1277, Feb. 2021.

[19] Y. Zou, K. T. Kim, X. Lin, and M. Chiang, "Minimizing age-of-information in heterogeneous multi-channel systems: A new partial-index approach," in *Proc. 22nd Int. Symp. Theory, Algorithmic Found., Protocol Design Mobile Netw. Mobile Comput.*, Jul. 2021, pp. 11–20.

[20] J. Pan, A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing age of information via scheduling over heterogeneous channels," in *Proc. 22nd Int. Symp. Theory, Algorithmic Found., Protocol Design Mobile Netw. Mobile Comput.*, Jul. 2021, pp. 111–120.

[21] B. Yin, S. Zhang, and Y. Cheng, "Application-oriented scheduling for optimizing the age of correlated information: A deep-reinforcement-learning-based approach," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8748–8759, Sep. 2020.

[22] Q. He, G. Dan, and V. Fodor, "Minimizing age of correlated information for wireless camera networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 547–552.

[23] B. Zhou and W. Saad, "On the age of information in Internet of Things systems with correlated devices," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[24] L. M. Hoang, J. Doncel, and M. Assaad, "Age-oriented scheduling of correlated sources in multi-server system," in *Proc. 17th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2021, pp. 1–6.

[25] J. Hribar, M. Costa, N. Kaminski, and L. A. DaSilva, "Using correlated information to extend device lifetime," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2439–2448, Apr. 2019.

[26] A. E. Kalor and P. Popovski, "Minimizing the age of information from sensors with common observations," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1390–1393, Oct. 2019.

[27] A. E. Kalør and P. Popovski, "Timely monitoring of dynamic sources with observations from multiple wireless sensors," 2020, *arXiv:2012.12179*.

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[29] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, pp. 287–298, Jan. 1988.

[30] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, Nov. 2010.

[31] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, p. 122, 2012.

[32] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2013, pp. 151–159.

[33] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, "Combinatorial multi-armed bandit with general reward functions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1651–1659.

[34] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2014, pp. 521–529.

[35] S. Paladino, F. Trovo, M. Restelli, and N. Gatti, "Unimodal Thompson sampling for graph-structured arms," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–7.

[36] J. Tong, S. Lai, L. Fu, and Z. Han, "Optimal frequency and rate selection using unimodal objective based Thompson sampling algorithm," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–7.

[37] S. Buccapatnam, F. Liu, A. Eryilmaz, and N. B. Shroff, "Reward maximization under uncertainty: Leveraging side-observations on networks," *J. Mach. Learn. Res.*, vol. 18, pp. 1–34, Apr. 2018.

[38] I. Kuzminykh, A. Carlsson, M. Yevdokymenko, and V. Sokolov, "Investigation of the IoT device lifetime with secure data transmission," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Cham, Switzerland: Springer, 2019, pp. 16–27.

[39] J. Wang, Y. Liu, and S. K. Das, "Energy-efficient data gathering in wireless sensor networks with asynchronous sampling," *ACM Trans. Sensor Netw.*, vol. 6, no. 3, pp. 1–37, Jun. 2010.

[40] H. Sharma, R. Jain, and A. Gupta, "An empirical relative value learning algorithm for non-parametric MDPs with continuous state space," in *Proc. 18th Eur. Control Conf. (ECC)*, Jun. 2019, pp. 1368–1373.

[41] M. Hauskrecht and B. Kveton, "Linear program approximations for factored continuous-state Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2003, pp. 895–902.

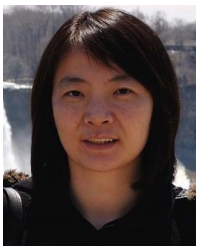
[42] L. Li and M. L. Littman, "Lazy approximation for solving continuous finite-horizon MDPs," in *Proc. AAAI*, Mar. 2005, pp. 1175–1180.

[43] L. B. Jackson, *Digital Filters and Signal Processing*. Boston, MA, USA: Kluwer, 1989.

[44] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge, Univ. Press, 2004.

[45] *Study on Channel Model for Frequencies From 0.5 to 100 GHz (Release 134)*, document GT 38.901, 3GPP, Jan. 2018.

Jingwen Tong (Student Member, IEEE) received the B.E. degree in electrical engineering from China Jiliang University, Hangzhou, China, in 2015, and the M.S. degree in electrical engineering from Ningbo University, Ningbo, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Communication Engineering, Xiamen University, Xiamen, China. He was a Visiting Ph.D. Student with the University of Houston, Houston, USA, from 2019 to 2020. His main research interests focus on multi-armed bandit, game theory, and wireless communications.



Liqun Fu (Senior Member, IEEE) received the Ph.D. degree in information engineering from The Chinese University of Hong Kong in 2010.

She was a Post-Doctoral Research Fellow with the Institute of Network Coding, The Chinese University of Hong Kong, from 2011 to 2013; and the ACCESS Linnaeus Centre, KTH Royal Institute of Technology, from 2013 to 2015. She was an Assistant Professor with ShanghaiTech University from 2015 to 2016. She is currently a Full Professor with the School of Informatics, Xiamen University, China. Her research interests are mainly in communication theory, optimization theory, and learning theory, with applications in wireless networks. She is on the Editorial Board of IEEE ACCESS and the *Journal of Communications and Information Networks* (JCIN). She served as the Technical Program Co-Chair of the GCCCN Workshop of the IEEE INFOCOM 2014, the Publicity Co-Chair of the GSNC Workshop of the IEEE INFOCOM 2016, and the Web Chair of the IEEE WiOpt 2018. She also serves as a TPC Member for many leading conferences in communications and networking, such as the IEEE INFOCOM, ICC, and GLOBECOM.

Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University in 1997 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was a Research and Development Engineer with JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate with the University of Maryland. From 2006 to 2008, he was an Assistant Professor with Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department and the Computer Science Department, University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of communications systems (Best Paper Award in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in IEEE conferences. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018. He has been an AAAS Fellow since 2019. He has been an ACM Distinguished Member since 2019. He has been 1% highly-cited researcher since 2017 according to Web of Science. He is also the Winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."