Decentralized Multi-Agent Bandit Learning for Intelligent Internet of Things Systems

Qiuyu Leng*†‡, Shangshang Wang*, Xi Huang§, Ziyu Shao*, Yang Yang*
*School of Information Science and Technology, ShanghaiTech University
†Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
‡University of Chinese Academy of Sciences
§Shenzhen Institute of Artificial Intelligence and Robotics for Society
Email: *{lengqy, wangshsh2, shaozy, yangyang}@shanghaitech.edu.cn, §dr.huangxi@gmail.com

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771884

***Abstract*—In intelligent Internet of Things systems, data-hungry services are empowered by *data collection*, which is jointly accomplished by *edge servers* and *data-collecting sensors*. In this paper, we aim to achieve efficient data collection, *i.e.*, maximize data rates from sensors to servers while mitigating the impact of data heterogeneity for data collected from sensors. Considering geographically distributed servers and sensors, we study the problem from the perspective of *multi-agent multi-armed bandits*. The key ideas of our approach are to 1) establish associations between servers and sensors under unknown wireless dynamics (*i.e.*, channel state information) and selection fraction constraints; 2) utilize shared information via pairwise communication between servers to mitigate biased observations for data rates. To this end, we propose a scheme that leverages online learning to reduce uncertainties in wireless dynamics and online control to mitigate the impact of data heterogeneity. Based on an effective integration of bandit learning methods under pairwise communication and Lyapunov optimization techniques, we present a novel *Decentralized sErver-Sensor association scheme with Multi-Agent learning under pairwise communication* (*DESMA*). Our theoretical analysis demonstrates that DESMA achieves a tunable trade-off between maximizing data rate and mitigating the impact of data heterogeneity.**

***Index Terms*—Intelligent Internet of Things systems, data heterogeneity, multi-agent bandit learning.**

I. Introduction

Recent advances in edge computing and machine learning techniques have given rise to intelligent Internet of Things (IoT) systems [1]. Intelligent IoT systems deliver intelligent services for data-hungry tasks (*e.g.*, neural network training [2]), which are empowered by *data collection*. Accordingly, data collection is jointly achieved by *edge servers* and *data-collecting sensors* in intelligent IoT systems.

To achieve efficient data collection, edge servers are employed to effectively gather information from data-collecting sensors via wireless channels [3]. It demands an optimization process that maximizes data rates from sensors to servers. However, such a process faces the concern of *geographically distributed* servers and sensors. Specifically, due to non-neglectable communication overhead, it is highly non-trivial to simultaneously execute centralized decisions for servers. A promising solution is to conduct decentralized decision-making processes, *i.e.*, each server dynamically associates data-collecting sensors individually (*a.k.a. decentralized server-sensor association*).

Towards effective design for decentralized server-sensor association, there remain several challenges. First, wireless dynamics (*i.e.*, channel state information) are usually time-varying and unknown. Second, each server may obtain biased observations of data rates after data collection [4]. Third, data on sensors is naturally heterogeneous. Poorly-conducted data collection may result in low-quality data and further impact the performance of intelligence services.

Based on the above discussion, we cast decentralized server-sensor association as a *multi-agent multi-armed bandit* (MAMAB) problem. Specifically, we view edge servers as agents and data-collecting sensors as arms. However, classical methods from the bandit field may fail to handle the biased observations for data rates and mitigate the impact of data heterogeneity on sensors. Effectively solving the concerns demands *information sharing* and *online control*, respectively. On one hand, information sharing between agents should be ensured to mitigate biased observations for data rates. In this paper, we assume that agents communicate and share information in a *pairwise communication pattern* under resource limitations [5]. On the other hand, we need to incorporate online control to mitigate the impact of inherent data heterogeneity on distinct servers. Nevertheless, the integration of online learning and online control is highly non-trivial since 1) ineffective online learning caused by biased observations may misguide the control procedure towards sub-optimal decisions; 2) wrongly performed online control may incur inferior feedback and further disrupt learning efficiency subsequently.

In this paper, we propose a *Decentralized sErver-Sensor association scheme with Multi-Agent learning under pairwise communication* called *DESMA*. With an integration of online learning and online control, DESMA maximizes data rates over servers while mitigating the impact of data heterogeneity on sensors. The main contributions are summarized as follows:

- **Problem Formulation:** We formulate the server-sensor association from the perspective of MAMAB (Sections II and III), aiming to maximize data rate while mitigating the impact of data heterogeneity.
- **Algorithm Design:** With an integration of bandit learning methods under pairwise communication and Lyapunov optimization techniques, we propose a novel scheme called DESMA to solve the formulated problem (Section III).
- **Theoretical Analysis:** Our analysis shows that each agent achieves a time-averaged regret $O(1/V + \sqrt{MK \log T/T})$

under DESMA, where T is the time horizon, M is the number of servers, K is the number of sensors, and V is a positive tunable parameter (Section IV).

- **Simulation Results:** Our results verify the effectiveness of DESMA in terms of maximizing data rates and show that DESMA and its variants achieve up to 10.45% higher data rates than the baseline method (Section V).

II. System Model and Problem Formulation

A. Basic Model

We consider an intelligent Internet of Things (IoT) system which comprises numerous geographically distributed edge servers and data-collecting sensors [6]. Suppose that there are M servers and K sensors whose index sets are denoted as $\mathcal{M} \triangleq \{1, \cdots, M\}$ and $\mathcal{K} \triangleq \{1, \cdots, K\}$, respectively. The system operates in a time-slotted fashion with time horizon T, indexed by $t \in \{0, 1, \cdots, T-1\}$.

B. Association Decisions

During each time slot, each edge server associates with *one* of the data-collecting sensors while a sensor can be associated with multiple servers at the same time. We assume that there is no collision among servers when simultaneously associating with the same sensor.

We define the association decision of server m during time slot t as $a_m(t) \in \mathcal{K}$, and summarize the association decisions of all servers by set $\mathcal{A}(t) \triangleq \{a_m(t) | \forall m \in \mathcal{M}\}$. We also denote $d_{m,k}(t) \triangleq \mathbb{1}\{a_m(t) = k\}$ as the decision indicator for whether the server m associates with the sensor k. Since each edge server can associate with only one sensor, we have $\sum_{k \in \mathcal{K}} d_{m,k}(t) = 1$.

Due to the dynamic channel conditions between servers and sensors, data rates are usually time-varying and unknown before data collection. When server m receives data from sensor k, we denote random variable $U_{m,k}(t)$ as the *true* data rate during data transmission process. We assume that $U_{m,k}(t)$ is independent and identically distributed (*i.i.d.*) across time slots and independent over edge servers. It varies in the interval $[u_{\min}, u_{\max}]$ with an *unknown* mean μ_k.

In practice, each server m may only obtain a *biased* observation of the true data rate $U_{m,k}(t)$ [4]. We denote such a data rate under biased observation (short as *observed data rate*) as random variable $U'_{m,k}(t)$ which we assume to be bounded within $[u_{\min}, u_{\max}]$ and *i.i.d.* with an *unknown* mean $\mu_{m,k}$. Note that the relationship between the true data rate $U_{m,k}(t)$ and the observed data rate $U'_{m,k}(t)$ is that $\mu_k = \frac{1}{M} \sum_{m=1}^{M} \mu_{m,k}$.

C. Pairwise Communication Setting

After associating with sensors, edge servers communicate with each other to share information about data rates during each time slot. The communication between servers is described by a *communication graph*. Specifically, the communication graph $G = (V, E)$ is a simple, undirected, and connected graph. Each vertex $v \in V$ represents an edge server $m \in \mathcal{M}$. Each edge $(i, j) \in E$ implies possible communication between servers i and j (denoted as neighboring servers), and we denote $\mathcal{N}_m$ as the set of server m's neighboring servers. Considering resource limitations on edge servers [7], we assume that servers communicate with others via a *pairwise* pattern [5]. Specifically, only *one pair* of neighboring servers (represented by vertexes v's in communication graph G) exchange information during each time slot. The communication among servers and sensors in the intelligent IoT system is demonstrated in Fig. 1.

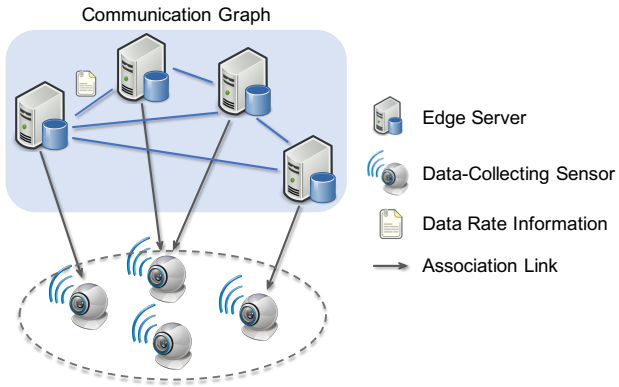


Fig. 1: Communication among edge servers and data-collecting sensors in intelligent IoT systems.

In time slot t, we activate exactly one pair of neighboring servers (i, j) in the edge set E of the communication graph G randomly and uniformly. The *communication matrix* W over communication graph G is denoted as

$$W \triangleq \frac{1}{|E|} \sum_{(i,j) \in E} \left(I_M - \frac{1}{2}(e_i - e_j)(e_i - e_j)^T \right),$$

where $|E|$ is the number of edges in communication graph G, $I_M \in \mathbb{R}^{M \times M}$ is an identity matrix, e_i denotes the N-dimensional unit vector with the ith entry equal to 1. Note that the communication matrix W is a positive semi-definite matrix whose largest eigenvalue equals 1 and its second largest eigenvalue is denoted by λ_2 ($\lambda_2 < 1$ since G is connected) [5].

D. Mitigating the Impact of Data Heterogeneity on Data-Collecting Sensors

In intelligent IoT systems, data on sensors is naturally heterogeneous. Poorly-conducted data collection may result in low-quality data and further impact the performance of intelligence services. For example, given that some classes account for a major proportion in the collected data (*i.e.*, class imbalance [8]), the neural network trained over such a dataset may be biased to those classes of samples while generalizing badly to other classes of small proportion.

To mitigate the impact of data heterogeneity, each server needs to guarantee adequate involvement of distinct sensors. Specifically, we need to ensure the minimum fraction of selection times (*a.k.a.* selection fraction) of sensor k for each server m. For each server m, we consider the following *selection fraction constraints* for each sensor k over time horizon T':

$$\liminf_{T' \to \infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}[d_{m,k}(t)] \geq c_{m,k}, \forall k \in \mathcal{K}, \tag{1}$$

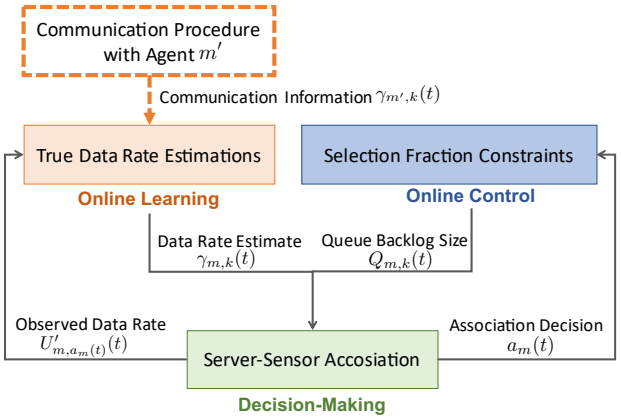


Fig. 2: An illustration of DESMA for agent m in time slot t.

where $c_{m,k} \in (0,1)$ is the minimum selection fraction of sensor k for server m. Note that $c_{m,k}$ is set based on the properties of collected data on sensors. For example, we set larger $c_{m,k}$ values for the sensors that have minority classes of collected data [8]. Intuitively, under such long-term time-averaged constraints, servers are expected to select each sensor with a selection fraction of at least $c_{m,k}$ in the long run.

E. Problem Formulation

In the intelligence IoT system, we aim to maximize total true data rates of all servers over time horizon T (short as *total data rate*) while mitigating the impact of data heterogeneity on sensors. Specifically, the formulated problem is as follows:

$$\begin{aligned} \underset{\{\mathcal{A}(t)\}_t}{\text{maximize}} \quad & \sum_{t=0}^{T-1} \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} \mathbb{E}[U_{m,k}(t) d_{m,k}(t)] \\ \text{subject to} \quad & \liminf_{T'\to\infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}[d_{m,k}(t)] \geq c_{m,k}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}. \end{aligned} \tag{2}$$

Note that the problem (2) is essentially a constrained stochastic optimization problem that involves the decision-making process of multiple edge servers under unknown wireless dynamics and selection fraction constraints.

III. Algorithm Design

Inspired by works [9] [10], we first reformulate problem (2) from the perspective of *multi-agent multi-armed bandits* (MAMAB) and then present our algorithm design.

A. Problem Reformulation

To reformulate problem (2) from the perspective of MAMAB, we view edge servers as agents and data-collecting sensors as arms. During each time slot t, each edge server (agent) m associates with (selects) one data-collecting sensor (arm) k. If arm k is selected by agent m, agent m would observe reward $R'_{m,k}(t) \triangleq U'_{m,k}(t)$. For each sensor k, we define its true reward as the true data rate, *i.e.*, $R_{m,k}(t) \triangleq U_{m,k}(t)$. With such a reformulation, problem (2) can be rewritten as a reward maximization problem as follows:

$$\begin{aligned} \underset{\{\mathcal{A}(t)\}_t}{\text{maximize}} \quad & \sum_{t=0}^{T-1} \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} \mathbb{E}[R_{m,k}(t) d_{m,k}(t)] \\ \text{subject to} \quad & \liminf_{T'\to\infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}[d_{m,k}(t)] \geq c_{m,k}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}. \end{aligned} \tag{3}$$

To characterize the *performance loss* due to decision-making process under unknown wireless dynamics and selection fraction constraints, for each agent m, we define the *time-averaged regret* over time horizon T as follows:

$$\text{Reg}_m(T) \triangleq R^*_m - \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k\in\mathcal{K}} \mathbb{E}[R_{m,k}(t) d_{m,k}(t)], \tag{4}$$

where R^*_m is the optimal expected reward for agent m.

Note that there are implicit couplings among agents since they need to communicate to mitigate biased observations and estimate the mean of true data rate μ_k for each sensor k. Moreover, the geographically distributed servers and sensors and their dynamic associations make problem (3) essentially a *decentralized decision-making problem under uncertainties subject to time-averaged constraints*.

Faced with such concerns, we adopt pairwise communications among agents (Section II-C) to share information and further mitigate biased observations. Under such a communication setting, the reward maximization problem (3) can be decoupled into the following sub-problem for each agent m [9]:

$$\begin{aligned} \underset{\{a_m(t)\}_t}{\text{minimize}} \quad & \text{Reg}_m(T) \\ \text{subject to} \quad & \liminf_{T'\to\infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}[d_{m,k}(t)] \geq c_{m,k}, \forall k \in \mathcal{K}. \end{aligned} \tag{5}$$

B. Decentralized Online Learning under Pairwise Communication

Online learning should be integrated into the decision-making process to cope with unknown wireless dynamics in intelligent IoT systems. The key issue is to handle the trade-off between *exploitation* (*i.e.*, leveraging current knowledge by associating with sensors with the empirically highest data rates) and *exploration* (*i.e.*, gaining new knowledge by associating with under-explored sensors in pursuit of high data rates in the long run) [11].

To address the exploitation-exploration dilemma, we utilize a variant of upper confidence bound (UCB) method [9] to estimate the true data rate, which is defined as follows:

$$\text{UCB}_{m,k}(t) \triangleq \gamma_{m,k}(t) + C_{m,k}(t), \tag{6}$$

where $\gamma_{m,k}(t)$ is an estimate of the true data rate from sensor k to server m, and $C_{m,k}(t)$ is the confidence radius.

Besides the unknown wireless dynamics, it is noteworthy that edge servers' inaccurate information of data rate may also lead to sub-optimal decision-making. Specifically, each server only possesses biased observations of data rates for sensors and the $\text{UCB}_{m,k}(t)$ is an inaccurate estimation of the true data rate μ_k.

To achieve effective online learning, servers require to share their learned knowledge with others to mitigate biased observations and learn the true data rate. In this paper, we adopt pairwise communications to share information among servers. During each time slot, *one pair* of neighboring servers in communication graph G is selected uniformly and randomly to exchange information, *i.e.*, estimates of the true data rate $\gamma_{m,k}(t)$, $\forall k \in \mathcal{K}$. Specifically, if agent m is selected to communicate with agent m' in time slot t, $\gamma_{m,k}(t)$ is updated as follows [9]:

$$\gamma_{m,k}(t+1) = \frac{\gamma_{m,k}(t) + \gamma_{m',k}(t)}{2} + \theta_{m,k}(t+1) - \theta_{m,k}(t). \quad (7)$$

Note that $\theta_{m,k}(t)$ is the empirical mean of $U'_{m,k}(t)$, which is computed as follows during each time slot t:

$$\theta_{m,k}(t+1) = \frac{1}{n_{m,k}(t+1)} \sum_{\tau=0}^{t+1} d_{m,k}(\tau) \cdot R'_{m,k}(\tau), \quad (8)$$

where $n_{m,k}(t+1)$ is the number of times arm k is selected by agent m till time slot $(t+1)$. If agent m is not selected to communicate with others, $\gamma_{m,k}(t)$ will be updated as follows:

$$\gamma_{m,k}(t+1) = \gamma_{m,k}(t) + \theta_{m,k}(t+1) - \theta_{m,k}(t). \quad (9)$$

Note that the confidence radius $C_{m,k}(t)$ in (6) is inversely proportional to $n_{m,k}(t)$. The fewer the times sensor k has been associated by server m, the higher the probability that it will be associated in the subsequent time slots. Such a term characterizes the uncertainty of data rate, and is defined as follows:

$$C_{m,k}(t) \triangleq u_0 \sqrt{\frac{2M \log t}{n_{m,k}(t)}} + \frac{64}{M^{17}}, \quad (10)$$

where $u_0 \triangleq u_{\max} - u_{\min}$ with $u_{\max}$ and $u_{\min}$ as the upper bound and lower bound of data rate, respectively.

***Remark* 1.** Compared with the confidence radius of classical UCB method [11], the additional coefficient $\sqrt{M}$ and additive term $64/M^{17}$ are introduced under the pairwise communication among agents [9].

C. Online Control for Selection Fraction Constraints

To mitigate the impact of data heterogeneity, we adopt *virtual queue* techniques [12] to handle the time-averaged constraints in (1). Specifically, for each agent m, we introduce a virtual queue for each arm $k \in \mathcal{K}$ which is denoted as $Q_{m,k}(t)$. The backlog size of each virtual queue is initialized as zero and then updated at the end of each time slot t as follows:

$$Q_{m,k}(t+1) \triangleq [Q_{m,k}(t) + c_{m,k} - d_{m,k}(t)]^+, \quad (11)$$

in which $[\cdot]^+ \triangleq \max\{\cdot, 0\}$. Note that if arm k is not selected by agent m in time slot t (*i.e.*, $d_{m,k}(t) = 0$), then the backlog size of virtual queue $Q_{m,k}(t+1)$ increases. Otherwise, it decreases.

Constraints in (1) are satisfied only when the queueing process $\{Q_{m,k}(t)\}_t$ for each agent $m \in \mathcal{M}$ is strongly stable [12], *i.e.*,

$$\limsup_{T' \to \infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}\left[\sum_{k \in \mathcal{K}} Q_{m,k}(t)\right] < \infty.$$

Algorithm 1 Decentralized sErver-Sensor association with Multi-Agent learning under pairwise communication (DESMA)

Input: The number of edge server M, the number of data-collecting sensor K, time-horizon T, and positive tunable parameter V.
Output: Association decision sequence $\{a_m(t)\}_t$ for each server m.
Initialize Each server associates with each sensor once, and observes data rate $U'_{m,k}(0)$, $k \in \mathcal{K}$. Set $n_{m,k}(0) \leftarrow 1$, $\theta_{m,k}(0) \leftarrow U'_{m,k}(0)$, $Q_{m,k}(0) \leftarrow 0$, $k \in \mathcal{K}$.
for $t \in \{0, 1, \cdots, T-1\}$ **do**
  $\mathcal{F}_m \leftarrow \varnothing$.
  $\tilde{n}_{m,k}(t) \leftarrow \max\{n_{m,k}(t), \tilde{n}_{m',k}(t), m' \in \mathcal{N}_m\}$, $\forall k \in \mathcal{K}$.
  Put k into set $\mathcal{F}_m$ **if** $n_{m,k}(t) < \tilde{n}_{m,k}(t) - M$, $\forall k \in \mathcal{K}$.
  %% *Association procedure for each server*
  Compute $\text{UCB}_{m,k}(t)$ according to (6), $\forall k \in \mathcal{K}$.
  if $\mathcal{F}_m = \varnothing$ **then**
    $a_m(t) \leftarrow \arg\max_{k \in \mathcal{K}} \Phi_{m,k}(t)$.
  else
    $a_m(t) \leftarrow \arg\max_{k \in \mathcal{F}_m} \Phi_{m,k}(t)$.
  end if
  Server m associates with sensor $a_m(t)$.
  Obtain observed data rate $U'_{m,a_m(t)}(t)$.
  $n_{m,a_m(t)}(t+1) \leftarrow n_{m,a_m(t)}(t) + 1$.
  %% *Communication procedure for each server*
  if server m is selected to communicate with server m' **then**
    server m sends $\gamma_{m,k}(t)$, $k \in \mathcal{K}$ to server m'.
    server m receives $\gamma_{m',k}(t)$, $k \in \mathcal{K}$ from server m'.
    Update $\gamma_{m,k}(t+1)$, $k \in \mathcal{K}$ according to (7).
  else
    Update $\gamma_{m,k}(t+1)$, $k \in \mathcal{K}$ according to (9).
  end if
  Update $Q_{m,k}(t+1)$, $k \in \mathcal{K}$ according to (11).
end for

To maintain the queue stability while maximizing the total data rate, we utilize Lyapunov optimization techniques to solve the problem (5), which can be approximately achieved when each agent m selects arm k that maximizes the following term:

$$\Phi_{m,k}(t) \triangleq Q_{m,k}(t) + V \cdot \min\{\text{UCB}_{m,k}(t), u_{\max}\}, \quad (12)$$

where V is a positive tunable parameter that balance the trade-off between minimizing regret and reducing backlog sizes of virtual queues, *i.e.*, maximizing data rate and mitigating the impact of data heterogeneity, respectively.

In summary, with an effective integration of online learning methods under pairwise communication and Lyapunov optimization techniques, we propose an online scheme called *Decentralized sErver-Sensor association with Multi-Agent learning under pairwise communication (DESMA)*. The pseudocode of DESMA for each agent m is demonstrated in Algorithm 1. The design of DESMA is shown in Fig. 2. The computational complexity of DESMA for each agent during each time slot is $O(K)$.

IV. Theoretical Analysis

The selection fraction vector $\boldsymbol{c}_m = (c_{m,1}, \ldots, c_{m,K})$ for each agent m is *feasible* if there exist some policies under which time-averaged constraints in (1) are satisfied. We define the set of all feasible selection fraction vectors for agent m as the *maximal feasibility region* $\mathcal{C}_m$. Based on such definitions, we have the following theorems. The detailed proofs of Theorems 1 and 2 are delegated to our technical report [13].

A. Selection Fraction Constraints

***Theorem* 1.** For each agent m, if $\boldsymbol{c}_m = (c_{m,1}, \ldots, c_{m,K})$ lies in the interior of $\mathcal{C}_m$, under DESMA, virtual queues satisfy:

$$\limsup_{T' \to \infty} \frac{1}{T'} \sum_{t=0}^{T'-1} \mathbb{E}\Big[\sum_{k \in \mathcal{K}} Q_{m,k}(t) \Big] \leq \frac{B + V u_{\max}}{\epsilon}, \qquad (13)$$

where ϵ is a positive constant which satisfies that $(\boldsymbol{c}_m - \epsilon \mathbf{1})$ is also an interior point of $\mathcal{C}_m$; V is a positive tunable parameter; $B = \sum_{k \in \mathcal{K}} \max\{c_{m,k}^2, (c_{m,k} - 1)^2\}/2$.

***Remark* 2.** Theorem 1 demonstrates that DESMA achieves the strong queue stability given a finite value of parameter V, which means that selection fraction constraints (1) are guaranteed [12]. Moreover, the time-averaged queue backlog size increases as the value of parameter V increases.

B. Bound of Time-Averaged Regret for Each Agent

***Theorem* 2.** If $\boldsymbol{c}_m = (c_{m,1}, \ldots, c_{m,K})$ lies in the interior of $\mathcal{C}_m$, each agent m achieves a time-averaged regret bound under DESMA as follows:

$$\mathrm{Reg}_m(T) \leq \frac{B}{V} + 4u_0\sqrt{2MK}\sqrt{\frac{\log T}{T}} + \frac{u_0}{T} \cdot \Bigg\{ K + 2K \cdot \max\Big\{L, (3K+1)M\Big\} + \frac{\pi^2 K}{3} + \frac{2K}{(1-\lambda_2^{1/3})(1-\lambda_2^{1/12})} \Bigg\},$$

where M is the number of edge servers (agents); K is the number of sensors (arms); $u_0 = u_{\max} - u_{\min}$; λ_2 is the second largest eigenvalue of the communication matrix W over communication graph G [9]; L is the minimal value that satisfies $\lambda_2^{t/6}/(1-\lambda_2^{1/3}) < (Mt)^{-1}$, $\forall\, t \geq L$.

***Remark* 3.** In classical communication setting (*i.e.*, agents only communicate to achieve consensus and do not incorporate online learning and online control), the smaller the value of λ_2, the faster the algorithm converges [5]. Note that Theorem 2 shows similar results: the smaller the value of λ_2, the lower regrets incurred by each agent (during online learning), which implies the faster DESMA learns the true data rates.

***Remark* 4.** Theorem 2 shows that the time-averaged regret for each agent is $O(1/V + \sqrt{MK \log T/T})$ with a tunable parameter V. Note that $O(1/V)$ term is mainly incurred during the online control procedure, mostly attributed to balancing the trade-off between maximizing data rate and mitigating the impact of data heterogeneity. The other term $O(\sqrt{MK \log T/T})$ is accumulated during the online learning procedure, which scales sub-linearly with the values of M and K. Intuitively, as either the number of servers or sensors increases, it will take a longer time for agents to learn the unknown wireless dynamics and mitigate the biased observations, hence incurring more regrets during the server-sensor association procedure.

V. Simulation Results

A. Simulation Settings

Basic Settings. We consider an intelligent IoT system with five edge servers ($M = 5$) and eight data-collecting sensors ($K = 8$). The time horizon T is set as 10^5. We consider a complete communication graph and the activated pair of servers that communicate with each other is selected randomly and uniformly. The communication matrix W is

$$\begin{bmatrix} 0.80 & 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.80 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.80 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.80 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.80 \end{bmatrix},$$

and its second largest eigenvalue is 0.75.

For sensor k and server m, the mean of observed data rate $\mu_{m,k}$ (in the unit of *bit/s*) is sampled from a truncated normal distribution with mean μ'_k and variance σ_k ($\sigma_k = 500$), and lies within the interval $[0, 6000]$. Note that μ'_k is sampled from a uniform distribution over the interval $[980, 5470]$, following the settings in [14]. The observed data rate $U'_{m,k}(t)$ is sampled from a truncated normal distribution with mean $\mu_{m,k}$ and variance $\sigma_{m,k}$ ($\sigma_{m,k}$=500), and lies within the interval $[0, 6000]$. According to Section II-B, the mean of true data rate μ_k is set as $\mu_k = \frac{1}{M}\sum_{m=1}^{M} \mu_{m,k}$. The true data rate $U_{m,k}(t)$ is sampled from a truncated normal distribution with mean μ_k and variance σ_k ($\sigma_k = 500$), and lies within the interval $[0, 6000]$.

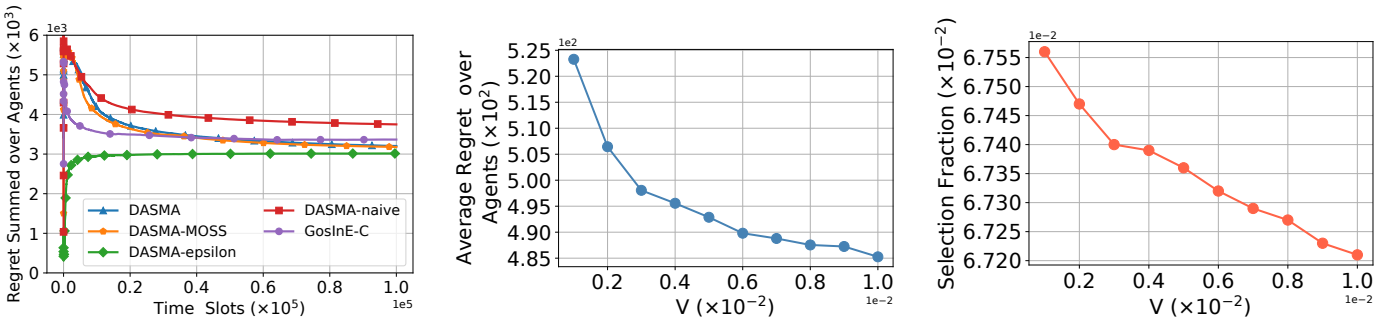
For selection fraction vector $\boldsymbol{c}_m = (c_{m,1}, \ldots, c_{m,K})$, we set $c_{m,k} = 0.067$ for all $m \in \mathcal{M}$, $k \in \mathcal{K}$. Note that we set the value of V to vary within $[0.001, 0.01]$ which makes magnitudes of $Q_{m,k}(t)$ and $\min\{\mathrm{UCB}_{m,k}(t), u_{\max}\}$ in (12) comparable, otherwise either term would dominate and result in trivial comparison of two terms.

Variants of DESMA. In addition to DESMA, we propose two variants by replacing the online learning procedure with other bandit learning methods.

- DESMA-MOSS: This variant follows the same decision making procedure as DESMA, except that (10) is replaced by a different term $C_{m,k}^{\mathrm{MOSS}}(t)$ [15] as follows:

$$C_{m,k}^{\mathrm{MOSS}}(t) \triangleq u_0 \sqrt{\frac{8M}{n_{m,k}(t-1)} \log^+\left(\frac{T}{K n_{m,k}(t-1)}\right)} + \frac{64}{M^{17}},$$

  where $\log^+(\cdot) \triangleq \log(\max\{1, \cdot\})$.
- DESMA-epsilon: During each time slot, each edge server associates with the sensor which has the empirically highest estimate of data rate with probability $(1-\epsilon)$; otherwise, each server uniformly and randomly selects and associates with one of the sensors. We set $\epsilon = 0.05$ in our simulations.
- DESMA-naive: This variant simply disables the communication procedure (lines 18-22 in Algorithm 1) in DESMA and does not mitigate the biased observations.



(a) Regrets under DESMA and Other Methods ($V = 0.01$).

(b) Regrets under Different Values of V.

(c) Selection Fractions for the Least Selected Sensor under Different Values of V.

Fig. 3: Performances of DESMA and Other Methods.

Baseline Method. We consider a related method called GosInE [16], which is a decentralized MAMAB algorithm under which agents exchange optimal arm index through pairwise communication. Different from DESMA, GosInE does not consider the biased observations and the impact of data heterogeneity. We modify GosInE with additional selection fraction consideration (denoted as *GosInE-C*) and employ it as the baseline method.

B. Performance Evaluation

Comparison with Other Methods. In Fig. 3(a), we evaluate the *total regret summed over agents* under DESMA and its variants, and the baseline method. Among these algorithms, DESMA-naive incurs the worst performance with total regret 3749.22 as it disables the communication procedure among agents. Compared with DESMA and its variants, GosInE-C performs the worst with total regret 3366.21 and converges to suboptimal results as it does not consider the biased observations. DESMA, DESMA-MOSS, and DESMA-epsilon achieve comparably performances with total regret 3203.19, 3179.32, 3014.55, respectively. Compared with the baseline method GosInE-C, DESMA and its variants achieve up to 10.45% reduction in regrets.

Effect of Parameter V. In Fig. 3(b) and Fig. 3(c), we investigate the impact of V on the average regret over agents and selection fraction for the least selected sensor, respectively. Note that the larger the value of V, the lower the total regret and the lower the selection fraction. Such results demonstrate the trade-off between maximizing data rate and mitigating the impact of data heterogeneity as suggested by Theorem 2.

VI. Conclusion

This paper studied efficient data collection in intelligent Internet of Things systems. We reformulated decentralized server-sensor association from the perspective of multi-agent multi-armed bandits. With an effective integration of online learning methods under pairwise communication and Lyapunov optimizations techniques, we proposed DESMA to maximize data rates while mitigating the impact of data heterogeneity. The theoretical analysis demonstrated that DESMA achieves a tunable trade-off between maximizing total data rate and mitigating the impact of data heterogeneity. Our simulation results showed that DESMA and its variants achieve up to 10.45% higher data rates than the baseline method.

References

[1] H. Li, K. Ota, and M. Dong, "Learning iot in edge: Deep learning for the internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of CVPR*, 2009.

[3] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Data collection and wireless communication in internet of things (iot) using economic analysis and pricing models: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2546–2590, 2016.

[4] A. Grammenos, C. Mascolo, and J. Crowcroft, "You are sensing, but are you biased? a user unaided sensor calibration approach for mobile sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–26, 2018.

[5] Y. Liu, J. Liu, and T. Basar, "Differentially private gossip gradient descent," in *Proceedings of CDC*, 2018.

[6] F. Hu and Q. Hao, *Intelligent sensor networks: the integration of sensor networks, signal processing and machine learning*. CRC Press, 2012.

[7] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.

[8] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Addressing class imbalance in federated learning," in *Proceedings of AAAI*, 2021.

[9] Z. Zhu, J. Zhu, J. Liu, and Y. Liu, "Federated bandit: A gossiping approach," in *Proceedings of SIGMETRICS*, 2021.

[10] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1799–1813, 2019.

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[12] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[13] Q. Leng, S. Wang, X. Huang, Z. Shao, and Y. Yang, "Decentralized multi-agent bandit learning for intelligent internet of things systems," ShangheiTech University, Tech. Rep., 2021. [Online]. Available: http://faculty.sist.shanghaitech.edu.cn/faculty/shaozy/DESMA.pdf

[14] D.-Y. Kim, S. Kim, H. Hassan, and J. H. Park, "Adaptive data rate control in low power wide area networks for long range iot services," *Journal of Computational Science*, vol. 22, pp. 171–178, 2017.

[15] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits." in *Proceedings of COLT*, 2009.

[16] R. Chawla, A. Sankararaman, A. Ganesh, and S. Shakkottai, "The gossiping insert-eliminate algorithm for multi-agent bandits," in *Proceedings of AISTATS*, 2020.