IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 7, 2017, accepted January 21, 2018, date of publication February 2, 2018, date of current version March 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2801350

Relay Selection for Underwater Acoustic Sensor Networks: A Multi-User Multi-Armed Bandit Formulation

XINBIN LI[1], (Member, IEEE), JIAJIA LIU[1], LEI YAN[1], (Student Member, IEEE), SONG HAN[1], AND XINPING GUAN[2], (Fellow, IEEE)

[1]Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao 066004, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200030, China

Corresponding author: Lei Yan (lyan@stumail.ysu.edu.cn)

This work was supported in part by the Key Project of National Nature Science Foundation China under Grant 61633017, in part by the National Nature Science Foundation China under Grant 61571387, and in part by the Graduate Student Innovation Project of Hebei Province under Grant CXZZSS2017049.

ABSTRACT Multi-user cooperative transmission is an attractive architecture for underwater acoustic sensor networks (UASNs). Cooperative transmission depends on careful allocations of resources such as relay selection, but traditional relay selection requires precise measurements of channel state information, which is infeasible for multi-user cooperative transmission due to the unique features and hardware restrictions of UASNs. In this paper, we model multi-user relay selection under a multiuser multi-armed bandit (MU-MAB) framework, whereby users are not provided any prior knowledge about underwater acoustic channel conditions. We first exploit a novel MU-MAB algorithm, DSMU-MAB, for relay selection, assuming that the reward distributions are initially unknown but remain constant. Second, we consider an evolving environment in which the reward distributions undergo changes in time, and DSMU-rMAB, a derivative of DSMU-MAB, is proposed, which can be robust to abrupt changes in underwater communication environments. The proposed algorithms not only help sources find the suitable relays to achieve a high quality transmission and avoid collisions among users but also reduce the mass of information exchanged among users. We established the effectiveness of our proposed algorithms using theoretical and numerical analyses.

INDEX TERMS Multi-user multi-armed bandit (MU-MAB), stable matching, distributed relay selection, UASNs.

I. INTRODUCTION

In recent years, Underwater Acoustic Sensor Networks (UASNs) have attracted growing interest, due to their use in on-going support applications for mineral exploitation, environmental monitoring, disaster prevention, military surveillance and safety systems [1]–[3]. To implement spatial diversity and overcome the effects of fading, cooperative transmission has been introduced into UASNs [4]–[6], in which the data collected by sensors will be relayed by wireless acoustic nodes to cooperatively fulfil tasks. With the increase in the number of sensors deployed underwater over the years, multiple source nodes jointly using multiple relay nodes have become more common in order to improve cooperative transmission utilization. Relay selection in multi-user UASNs is a problem that should not be neglected; however, it is a unique challenge to enable multiple source nodes to select and share a common set of relays in an efficient and fair way.

In conventional radio frequency (RF) or UASN single-source scenarios, most relay selection architectures in the current literature are designed based on full instantaneous CSI [7] or statistical CSI [8], [9] feedback to achieve adaptive decision-making. In fact, CSI-based relay selection does not perform well in UASNs, especially in multi-user scenarios, for the following reasons.

1) Doosti-Aref and Ebrahimzadeh [7], considered relay selection with perfect channel state information (CSI) at the source. However, due to the harsh underwater environments and propagation delays (five orders of magnitude higher than in RF terrestrial channels), the CSI at the transmitter was actually imperfect.
2) In some of the literature, Wei and Kim [8] and Luo *et al.* [9] have considered relay selection based on prior statistical CSI. However, in consideration of highly dynamic changes in shallow seabeds or transient acoustic channel access to other artificial acoustic systems,

a relay selection policy based on prior statistical CSI may not be robust to abrupt changes in communication conditions.

3) In multi-user scenarios, along with the increasing number of source nodes, CSI information feedback will increase, which yields excessive overhead and computational costs.

In this paper, we present a novel multi-armed bandit (MAB) decision-making learning framework to solve the relay selection problem without knowledge of the channel at the source. MAB has been widely applied in website optimization [10]–[12] and optimal control strategies for robots [13], [14]. Recently, MAB has been widely used to address wireless communications and networking decision-making problems [15]–[17]. In a stochastic MAB problem, given a set of arms (actions), an arm is played (selected) at each trial and receives a reward drawn from the reward generating process of that arm. A stochastic reward with an unknown mean is associated with each arm, and upon pulling a single arm, the player receives an instantaneous reward. The player decides which arm to pull in a sequence of trials to maximize its accumulated reward over the long run. Every MAB model is a class of sequential decision-making problems under strictly limited prior information and feedback [16], and by employing this learning framework, we can perform relay selection based on instantaneous rewards rather than feeding knowledge regarding CSI back to the source.

In fact, multi-user relay selection problems are not simple superpositions of single-source relay selections; in a multiuser scenario, a relay selection scheme is needed to determine how relays are assigned [18] in order to maximize whole network performance. An effective allocation mechanism is an indispensable part of multi-user relay selection, the aim of which is to reduce collisions among users that occur when more than one source node user accesses the same relay simultaneously. In addition, in consideration of the reliability of data transmission and energy constraints (nodes equipped with batteries cannot be recharged), distributed algorithms are more applicable to UASNs, which are robust to transient losses of connectivity and limited information exchange. Therefore, we present two new algorithms based on MAB to further improve the performance of multi-user relay selection for UASNs.

1) We modified the current MAB framework using stable matching theory and a back-off timer and present a distributed multi-user multi-armed bandit (MU-MAB) relay selection algorithm, DSMU-MAB. Stable matching theory is a well-known, Nobel-prize winning framework that was introduced in a paper by Gale and Shapley [19]. Under this kind of one-to-one selection scheme, multisource access collisions can be avoided. Moreover, a back-off timer was designed to reduce information exchange among users.
2) In consideration of highly dynamic changes in shallow seabeds or transient acoustic channel access from other artificial acoustic systems such as UASNs and sonar users or natural acoustic systems such as marine mammals [10], a distributed stable matching multi-user **robust** MAB algorithm, DSMU-*r*MAB, is proposed to overcome abrupt changes in communication conditions under a non-stationary MAB setting where the reward distributions undergo changes in time.

Specifically, we provide the following contributions in this work.

- To the best of our knowledge, we present herein the first analysis of a relay selection problem in a more practical scenario with multiple sources and multiple relays for UASNs. In addition, we present a learning framework to learn multi-user relay selection as a MAB problem without any prior knowledge regarding the nature of the environment (i.e., instantaneous full CSI or knowledge of channel statistics).
- To reduce the number of computations and solve the conflict problem in a distributed way, we present a novel MU-MAB algorithm, DSMU-MAB, by employing stable matching theory and a back-off timer to realize stable relay selection in which collisions are eliminated to ensure efficient communication and to avoid masses of information exchange. We also present DSMU-*r*MAB, a derivative of DSMU-MAB, to overcome abrupt changes in communication environments, which has not been mentioned in other current, related works on MU-MAB.
- We present the regret upper bound of DSMU-MAB and DSMU-*r*MAB. Our simulation results show that our design employing small amounts of information exchange can achieve comparable performance to that of existing MU-MAB algorithms, and DSMU-rMAB can be robust to abrupt changes in underwater communication environments.

The rest of this paper is organized as follows. Section II presents the background on MAB theory along with related work on MU-MAB. Section III introduces the system model. Section IV describes in detail how we map the our problem into the learning framework and derive DSMU-MAB algorithm for relay selection. In Section V, we consider an evolving environment where the reward distributions undergo changes in time, DSMU-*r*MAB, a derivative of DSMU-MAB is proposed. In Section VI, we present the simulation results analysis to validate the performance of our DSMU-MAB and DSMU-*r*MAB algorithm. At last, we summarize the paper in Section VII.

II. RELATED WORK

MAB is a fundamental reinforcement learning framework for learning unknown parameters and has been widely used in wireless communications and networking. Zhao and Gai utilized MAB to formulate an opportunistic spectrum access problem [20]–[23] in cognitive radio scenarios. Maghsudi applied MAB to model efficient resource allocation in

[24] and [25]. Nikfar and Vinck [26], applied MAB to model efficient relay selection for cooperative power line communication. Shankar and Chitre [27] and Jayasuriya [28], first applied MAB to tune underwater physical link parameters. Furthermore, in [29], we first presented single-user relay selection based on MAB for underwater communication networks.

Several results from the MAB problem will be used and generalized to study our problem. Currently, much of the existing literature, including [20], [23], and [30], has studied the MU-MAB problem. Because there are multiple users in the systems, it is necessary to find allocation schemes to avoid collisions among users. Reference [30] presented the Hungarian MU-MAB policy for the problem of learning combinatorial matchings of users to resources. Reference [23] presented a novel policy, matching learning with polynomial storage (MLPS), that uses only polynomial storage and computation time at each decision period. A key subroutine of the MLPS policy involves solving a combinatorial optimization problem pertaining to weighted matchings with polynomial complexity at each step. Gai and Krishnamachari [23], [30], considered a combinatorial bandit framework, and the proposed schemes required a centralized coordinator, a large amount of information exchange and coordination among the users and mass information exchanges regarding the measured throughput on each user, which also led to large communication overheads. In [20], a distributed fair access scheme (DLF) was proposed to take into account collisions among users. Although that scheme can effectively avoid collisions among users, those distributed MU-MAB algorithms paid more attention to multiuser resource allocation in symmetric cases, such that each user obtains the same reward on a given arm. Obviously, due to geographic dispersion or underwater obstacles such as fish schools [9], users always obtain different rewards on a given relay, and asymmetrical cases are more common in multiuser UASN relay selection. Huang *et al.* [31], presented an OLGS algorithm and applied stable matching theory to achieve a distributed adaptive distributed channel allocation in an asymmetrical opportunistic spectrum access system based on empirical evidence of efficiency but without theoretical analysis. Moreover, no one has considered non-stationary settings where the reward distributions undergo changes in time in the MU-MAB problem presented above.

In our work, we neither appoint a coordinator nor limit our scheme to symmetrical cases. To overcome the drawbacks of the allocations described above, we present a DSMU-MAB algorithm based on stable matching and design back-off timers to reduce information exchange. The MU-MAB algorithm can be used in symmetrical cases and asymmetrical cases in a distributed way. In addition, DSMU-rMAB is presented to against abrupt changes in underwater communication environments. Detailed theoretical regret analyses of DSMU-MAB and DSMU-rMAB are presented in our work.

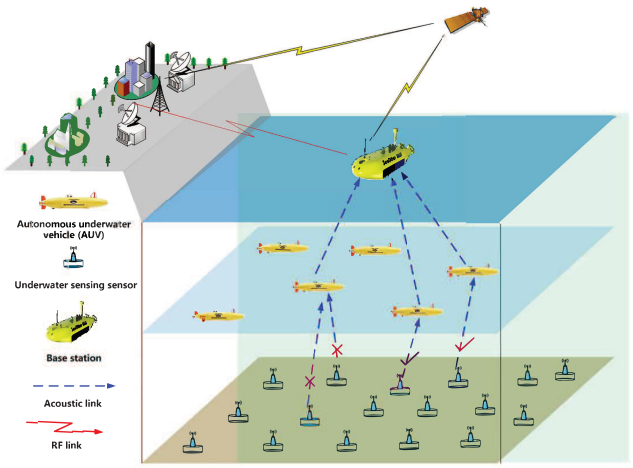


FIGURE 1. Cooperative transmission system model.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this paper, an UASNs cooperative transmission scenario is proposed,[1] as shown in Fig. 1. We consider a slotted UASN wherein each source node can access only one relay in each transmission slot. The system model consists of M source nodes equipped with acoustic modem, $K \geq M$ mobile autonomous underwater vehicles (AUVs) and a base station (BS) at the surface. The time is slotted, and we denote n as the total number of time slots t such that $1 \leq t \leq n$ is any arbitrary time slot. At time t, each source node can select a relay only based on its own observation histories under a decentralized policy and transmit its data to the BS assisted by the relay. If user $m \in \{1, \ldots, M\}$ at time t selects relay $k \in \{1, \ldots, K\}$, assuming no other conflicting users select that relay it gets an instantaneous reward $X_{m,k}(t)$.[2] Otherwise, if multiple users are selecting the same relay, then we assume that, due to collisions, none of the conflicting users derive any benefit. We assume that $X_{m,k}(t)$ follows some unknown i.i.d. process over time, with the only restriction that its distribution has a finite support. Without loss of generality, we normalize $X_{m,k}(t) \in [0, 1]$. The mean of random variable $X_{m,k}(t)$ is $\theta_{m,k} = \mathbb{E}_n[X_{m,k}(t)]$, which is unknown to the users and distinct from others. We denote the set of all these means as $\Theta = \{\theta_{m,k}, 1 \leq m \leq M, 1 \leq k \leq K\}$. In addition, the users need not to obtain CSI and any priori knowledge of the matrix of mean values, they only have to estimate and predict relay availability by exploring and learning. We denote kk^* as a set of M largest expected rewards for user-relay pairs.

The performance of a relay selection is evaluated by its regret value, which is defined as the difference between the expected reward that could be obtained by a genie that can

[1] Our scheme do not only used in this kind of vertical underwater acoustic links, for example it also can used for cooperative transmission among AUVs in horizontal links.

[2] The selection of reward metrics is dependent on the specific system implementation and based on the desired objective, for UASNs, we can denote throughput, delay, energy consumption, packet error ratio as the reward metrics.

pick the optimal arm at each time, and that obtained by the given policy π. We then can obtain the mathematical expression for the stationary regret after n time slots:

$$\mathfrak{R}_1^{\pi}(\Theta; n) = n \sum_{(m,k)\in kk^*} \theta_{m,k} - \mathbb{E}^{\pi}[\sum_{t=1}^{n} S_{\pi(t)}(t)] \quad (1)$$

$S_{\pi(t)}(t)$ is the sum of the actual reward obtained by all users at time under policy $\pi(t)$, which could be expressed as:

$$S_{\pi(t)}(t) = \sum_{k=1}^{K}\sum_{m=1}^{M} X_{m,k}(t) \times \mathbb{I}_{m,k}(t), \quad (2)$$

In formula (2), $\mathbb{I}_{m,k}(t)$ reflects the collision between source nodes in slot t, when source node m is the only one to select relay k, then $\mathbb{I}_{m,k}(t) = 1$, otherwise $\mathbb{I}_{m,k}(t) = 0$.

We then consider a non-stationary case, where the reward distributions undergo changes in time; in other words, $\theta_{m,k}$ may change over time. The regret defined in (1) is no longer appropriate for the time-invariant case. The mathematical expression for this non-stationary regret under policy π is:

$$\mathfrak{R}_2^{\pi}(\Theta(t); n) = \sum_{t=1}^{n} \sum_{(m,k)\in kk^*(t)} \theta_{m,k}(t) - \mathbb{E}^{\pi}[\sum_{t=1}^{n} S_{\pi(t)}(t)] \quad (3)$$

Table 1 presents a detailed list of the notation used throughout the paper.

TABLE 1. Notation.

K	the number of relays
M	the number of users
n	the total number of time slots
t	$1 \le t \le n$ is any arbitrary time slot
m, k	user-relay pair (user m, relay k)
kk	a matching contains M user-relay pairs
kk^*	a optimal expected matching contains M user-relay pairs
$X_{m,k}(t)$	an instantaneous reward that user m gets when m selects relay k at time t
$\theta_{m,k}$	the expectation of instantaneous reward of total n trials
$n_{m,k}$	number of times that relay k has been selected by user m up to the current time slot
$\hat{\theta}_{m,k}$	average of all the observed values of relay k by user m up to the current time slot
$\bar{n}_{m,k}$	the discounted number of times that relay k has been selected by user m up to the current time slot
$\hat{\theta}'_{m,k}$	the discounted empirical average of all the observed values of relay k by user m up to the current time slot
$b_{m,k}$	UCB index of user-relay pair at current time slot
$n_{i,j}^{kk}(t)$	$(i,j) \in kk(t)$, number of times that relay j has been selected by user i up to current time t
$n_{i,j}^{kk^*}(t)$	$(i,j) \in kk^*(t)$, number of times that relay j has been selected by user i up to current time t
$\bar{n}_{i,j}^{kk}(t)$	$(i,j) \in kk(t)$, the discounted number of times that relay j has been selected by user i up to current time t
$\bar{n}_{i,j}^{kk^*}(t)$	$(i,j) \in kk^*(t)$, the discounted number of times that relay j has been selected by user i up to current time t
$\bar{\bar{\theta}}_{kk(t)}(t)$	$\sum_{(i,j)\in kk(t)} \hat{\theta}_{i,j}(t)$
$\bar{\bar{\theta}}'_{kk(t)}(t)$	$\sum_{(i,j)\in kk(t)} \hat{\theta}'_{i,j}(t)$
$\Delta_{i,j}^{kk}(t)$	$\theta_{i,j}^{kk^*}(t) - \theta_{i,j}^{kk}(t)$
$\Delta_{min}^{i,j}$	$\min_{kk}\{\Delta_{i,j}^{kk} : (i,j) \in kk\}$
Δ_{min}	$\min_{kk} \Delta_{i,j}^{kk}$
Δ_{max}	$\max_{kk} \Delta_{i,j}^{kk}$

IV. DSMU-MAB: DISTRIBUTED STABLE MATCHING MULTI-USER MAB FOR RELAY SELECTION

In this section, we consider the case whereby no prior reward distribution knowledge is provided throughout the relay selection process, but the distributions are assumed to remain constant during all games. The multi-user relay selection algorithm DSMU-MAB is proposed; the algorithm has a self-learning ability that can be implemented well in complex underwater environments and is based on a modified current MAB framework with stable matching theory and a back-off timer.

A. FORMULATION OF MU-MAB

For relay selection in UASNs, rather than relying on the availability of full, instantaneous CSI, we need to predict channel quality using a learning algorithm. The learning mechanism aims to exploit all gathered information to evaluate the most promising relays. We now consider a stationary formulation of the MAB whereby the reward distributions are fixed. In this section, we suggest a simple learning mechanism referred to as UCB, which borrows from the MAB in [32] and its extended form [23]. To provide an optimistic evaluation of the relay's quality, the UCB algorithm associates an index called the UCB index to each user-relay pair. The computed index for each user-relay pair is then used as an estimate for the corresponding reward expectations and to select the user-relay pair with the highest index. Our work is influenced by the formulation in [23], wherein each arm corresponds to a matching of users to relays. The key idea behind this algorithm is to store and use observations for each user-relay pair rather than for each arm as a whole [23]. We provide the combinatorial UCB algorithm in terms of types of feedbacks in combinatorial bandits [33]; our work belongs to the semi-bandit type, in which the user observes only the outcomes of selected relays in one round. In each round, an arm is selected, and the outcomes of its related reward of user-relay pairs are observed, which aids the selection of arms in future rounds.

At each time t, after a user-relay pair $(m, k) \in kk(t)$ is selected, we obtain the observation of $X_{m,k}(t)$ for all $(m, k) \in kk(t)$ ($kk(t)$ is a set (super arm) that contains M user-relay pairs at time t). $(\hat{\theta}_{m,k})_{M\times K}$ and $(n_{m,k})_{M\times K}$ are then updated as follows:

$$\begin{cases} \hat{\theta}_{m,k}(t) = \dfrac{\hat{\theta}_{m,k}(t-1)n_{m,k} + X_{m,k}(t)}{n_{m,k}(t-1)+1} \\ n_{m,k}(t) = n_{m,k}(t-1) + 1 \end{cases} \quad (4)$$

Our proposed scheme of selecting a set containing M user-relay pairs that maximizes the expected reward is expressed as Algorithm 1.

We use two M by K matrices to store the information. One is $(\hat{\theta}_{m,k})_{M\times K}$, and the other is $(n_{m,k})_{M\times K}$. Both are calculated by formula (4) after we select a user-relay pair at each time slot. In the above algorithm, line 4 ensures that there will be

Algorithm 1 DSMU-MAB

// INITIALIZATION
for $t = 1 \rightarrow K$ **do**
  for $m = 1 \rightarrow M$ **do**
    Select relay k such that $k = ((m + t) mod K) + 1$;
    $\hat{\theta}_{m,k}(t) = X_{m,k}(t)$;
    $n_{m,k}(t) = 1$;
  end for
end for
// MAIN LOOP
while 1 **do**
  $t = K + 1$ do
  Run algorithm 2 to get a set contains M user-relay pairs that maximizes

$$\sum_{(p,q)\in kk(t)} \hat{\theta}_{p,q}(t) + \sqrt{\frac{2\ln t}{n_{p,q}(t)}} \tag{5}$$

  Update $(\hat{\theta}_{m,k})_{M\times K}$, $(n_{m,k})_{M\times K}$ accordingly.
end while

no collisions among users. Our scheme selects M user-relay pairs with the maximum value $b(kk)$ at each time slot after the initialization period, when each user-relay pair is chosen once.

B. STABLE MATCHING SCHEME

We will denote the UCB index of user m when using relay k by $b_{m,k}$ and define the UCB index matrix as $(B_{m,k})_{M\times K}$. At any given time the $M \times K$ user-relay pairs UCB index values are almost surely all different. To this end we need some definitions:

Definition 1: A matching between users and relays is a one-to-one[3] *function* $kk : [M] \longrightarrow [K]$ *where* $[M] = \{1, \ldots, M\}$.

We define the total UCB index of a matching kk by

$$b(kk) := \sum_{m=1}^{M} b_{m,kk_m(t)} \tag{6}$$

Indeed, there has been a recent surge in papers concerning possible applications of centralized optimization to solve assignment problems. Centralized optimization is a new mathematical tool for optimizing assignments in many emerging wireless systems. However, centralized optimizations often require global network information and centralized control, which thus yield significant overhead and computational complexity. Complexity can rapidly increase when dealing with combinatorial, integer programming problems [34]. The optimal centralized relay allocation problem is now formalized as follows:

$$kk_{opt} = argmax\{b(kk)|kk : [M] \xrightarrow{1-to-1} [K]\}. \tag{7}$$

[3]We adopt the "one-to-one matching" theory in [34], and in the relay selection setting, it means that source node m can be matched to at most one member of the opposite relay set [K]={1,...,K}.

The Hungarian centralized optimization scheme can provide optimal solutions, and its algorithmic implementations have matured over the past few years [35]. Optimizing relay allocations for underwater cooperative transmission using centralized optimization can result increased overhead due to information exchange and centralized computation. Here, we are interested in efficient distributed schemes that are suitable to UASNs.

To overcome the limitations of the Hungarian centralized optimization allocation described in the references cited above, we analyzed a distributed allocation scheme based on stable matching theory in an efficient and computationally inexpensive way. The relay allocation problem can be posed as a stable matching problem between relays and users. The main goal of matching is to optimally match relays and users, given their individual and learned information. Each source node user builds a ranking of the relays using a preference relation. Note that in this case, a preference can simply be defined in terms of a UCB index that predicts the higher throughput achieved by a certain user-relay matching. Now we can call a matching stable when no user-relay pairs prefer each other in comparison to their current matching. Hence, we obtain that in our case, stability is defined as follows:

Definition 2 [36]: A matching $S : [M] \longrightarrow [K]$ *is stable if for every* $m \in [M]$ *and* $k \in [K]$ *satisfying* $S(m) \neq k$ *if* $b_{m,S(m)} < b_{m,k}$ *then there exists some user* $m' \in [M]$ *such that* $S(m') = k$ *and* $b_{m',k} > b_{m,k}$.

The advantages of stable matching theory for relay allocation are as follows. (i) Because stable matching theory always specifies a stable one-to-one matching for any preference function, it can avoid multi-user contention under this interference model. (ii) Stable matching theory predicts unique stable matching when the entries of the preference matrix are all different. In [36], stable matching theory is used to obtain the only stable result proven stable. (iii) Stable matching theory allows each player (i.e., source node and relay) to define its individual utilities depending on its local information. We show that in our setting there is no significant overhead and that the computational complexity of the algorithm is greatly reduced.

To obtain a low-level information exchange implementation for the relay allocation, we would need to use a back-off timer, as shown in Fig. 2. We consider that all users choose the same back-off function, which is a monotonically decreasing function of their user-relay pair UCB index. At the beginning of each time slot after the initialization period, users calculate and set the back-off timer. Each user in the network calculates UCB index $b_{m,k}$ and maps it to a back-off time $\tau_{m,k}$ based on a predetermined common decreasing function $f(b_{m,k})$. Fig. 3 shows an example of such a back-off function. The first back-off that expires belongs to the user- relay pair that has the highest value in the matrix B. Source node m is to be allocated relay k. At each allocating period, source node m broadcasts an allocated message with format ⟨`index_m, index_k`⟩. This is equivalent to deleting a row and a column from matrix B, as implemented

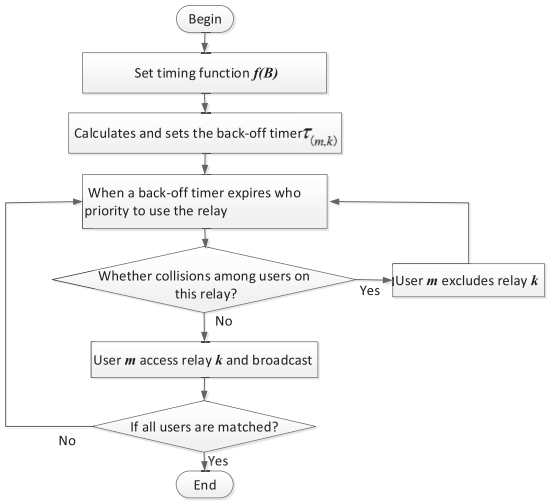


FIGURE 2. Flowchart of back-off timer.

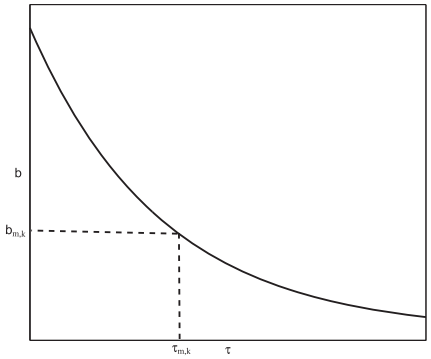


FIGURE 3. Example of a function that maps better UCB index to shorter time.

in Algorithm 2. The timer's value is adjusted according to the user-relay pair of the UCB index, and the user-relay pair of the larger UCB index is expected to end early. To clearly reflect that proposed scheme, we present the distributed relay allocation algorithm described above as Algorithm 2.

Theorem 1: The expected regret of DSMU-MAB is at most

$$[\frac{8MKlnn}{(\Delta_{min})^2} + M^2K(1+\frac{\pi^2}{3})]\Delta_{max} \tag{8}$$

Proof: Denote $C_{t,n_{i,j}} = \sqrt{\frac{(L+1)\ln t}{n_{i,j}}}$. We introduce $T_{i,j}(n)$ as a counter after the initialization period. It is updated in the following way:

At each time slot after the initialization period, when non-optimal set $kk(t)$ is selected at time t, there must be at least one user-relay pair $(i,j) \in kk(t)$, such that $(i,j) \notin kk^*(t)$. If there is only one such pair, $T_{i,j}(n)$ is increased by 1. If there are multiple such pairs, we arbitrarily pick one, say (p,q), and increment $T_{p,q}$ by 1.

Each time when a non-optimal set is picked, exactly one element in $(T_{i,j}(n))_{M\times K}$ is incremented by 1. This implies that the total number that we have played the non-optimal sets is equal to the summation of all counters in $(T_{i,j}(n))_{M\times K}$.

Algorithm 2 Stable Matching Subroutine

1: Input: M, K, $(\hat{\theta}_{m,k})_{M\times K}$, $(n_{m,k})_{M\times K}$
2: Output: kk, consisting of the M engaged user-relay pair
3: Calculate UCB index
4: Get the value of $(B_{m,k})_{M\times K}$
5: $(\tilde{B}_{m,k})_{M\times K} = (B_{m,k})_{M\times K}$
6: **for** $t = 1 \to M$ **do**
7: Get the maximum in the matrix $(\tilde{B}_{m,k})_{M\times K}$
8: // Assume that $\tau_{p,q}$ is the shortest time ($b_{p,q}$ is the maximum value in $(\tilde{B}_{m,k})_{M\times K}$, $p \in [1,M]$, $q \in [1,K]$)
9: update $(\tilde{B}_{m,k})_{M\times K}$;
10: $\forall 1 \le i \le M$, set $\tilde{B}_{i,q} = 0$;
11: $\forall 1 \le j \le K$, set $\tilde{B}_{p,j} = 0$;
12: **end for**
13: Terminate: A stable matching between users and relays

Therefore, we have:

$$\sum_{kk:\theta_{kk}<\theta_{kk*}} \mathbb{E}[T_{kk}(n)] = \sum_{i=1}^{M}\sum_{j=1}^{K} \mathbb{E}[T_{i,j}(n)] \tag{9}$$

Denote by $\mathbb{I}_{i,j}(n)$ the indicator function which defined to be 1 when the $T_{i,j}(n)$ is added by one at time n, and 0 when it is false. Let l be an arbitrary positive integer. Then:

$$\begin{aligned} T_{i,j}(n) &= \sum_{t=K+1}^{n} \{\mathbb{I}_{i,j}(t)\} \\ &\le l + \sum_{t=K+1}^{n} \{\mathbb{I}_{i,j}(t), T_{i,j}(t-1) \ge l\} \end{aligned} \tag{10}$$

When $\mathbb{I}_{i,j}(t) = 1$, there exists a user-relay pair $(i,j) \notin kk^*(t)$ such that a non-optimal set is picked. We denote this set as $kk(t)$ since at each time that $\mathbb{I}_{i,j}(t) = 1$. Then,

$$\begin{aligned} T_{i,j}(n) &\le l + \sum_{t=K+1}^{n} \{\hat{\bar{\theta}}_{kk^*(t-1)}(t-1) + C_{t-1,n_{kk^*(t-1)}(t-1)} \\ &\le \hat{\bar{\theta}}_{kk(t-1)}(t-1) + C_{t-1,n_{kk(t-1)}(t-1)}, T_{i,j}(t-1) \ge l\} \end{aligned} \tag{11}$$

Note that $\hat{\bar{\theta}}_{kk^*(t-1)}(t-1) + C_{t-1,n_{kk^*(t-1)}(t-1)} \le \hat{\bar{\theta}}_{kk(t-1)}(t-1) + C_{t-1,n_{kk(t-1)}(t-1)}$, for $j \epsilon [1,K]$, at least one of the following must hold:

$$\begin{aligned} \hat{\theta}^{kk^*}_{1,j}(t) + C_{t,n^{kk^*}_{1,j}(t)} &\le \hat{\theta}^{kk}_{1,j}(t) + C_{t,n^{kk}_{1,j}(t)}, \\ \hat{\theta}^{kk^*}_{2,j}(t) + C_{t,n^{kk^*}_{2,j}(t)} &\le \hat{\theta}^{kk}_{2,j}(t) + C_{t,n^{kk}_{2,j}(t)}, \\ &\vdots \\ \hat{\theta}^{kk^*}_{M,j}(t) + C_{t,n^{kk^*}_{M,j}(t)} &\le \hat{\theta}^{kk}_{M,j}(t) + C_{t,n^{kk}_{M,j}(t)}. \end{aligned}$$

Note that $T_{i,j}(t) \ge l$ implies,

$$n_{i,j}(t) \ge T_{i,j}(t) \ge l. \tag{12}$$

This means:

$$\begin{aligned}
T_{i,j}(n) &\leq l + M\sum_{t=K+1}^{n}\{\hat{\theta}_{i,j}^{kk^*}(t) + C_{t,n_{i,j}^{kk^*}(t)} \\
&\leq \hat{\theta}_{i,j}^{kk}(t) + C_{t,n_{i,j}^{kk}(t)}, T_{i,j}(t-1) \geq l\} \\
&\leq l + M\sum_{t=K}^{n}\{\min_{0<n_{i,j}^{kk^*}<t} \hat{\theta}_{i,j}^{kk^*}(t) + C_{t,n_{i,j}^{kk^*}(t)} \\
&\leq \max_{l<n_{i,j}^{kk}<t} \hat{\theta}_{i,j}^{kk}(t) + C_{t,n_{i,j}^{kk}(t)}\} \\
&\leq l + M\sum_{t=1}^{n}[\sum_{n_{i,j}^{kk^*}=1}^{t-1}\sum_{n_{i,j}^{kk}=l}^{t-1} \hat{\theta}_{i,j}^{kk^*}(t) + C_{t,n_{i,j}^{kk^*}(t)} \\
&\leq \hat{\theta}_{i,j}^{kk}(t) + C_{t,n_{i,j}^{kk}(t)}] \qquad (13)
\end{aligned}$$

Now observe that $\hat{\theta}_{i,j}^{kk^*}(t) + C_{t,n_{i,j}^{kk^*}(t)} \leq \hat{\theta}_{i,j}^{kk}(t) + C_{t,n_{i,j}^{kk}(t)}$ implies that at least one of the following must be true:

$$\hat{\theta}_{i,j}^{kk^*}(t) \leq \theta_{i,j}^{kk^*}(t) - C_{t,n_{i,j}^{kk^*}(t)} \qquad (14)$$

$$\hat{\theta}_{i,j}^{kk}(t) \leq \theta_{i,j}^{kk}(t) - C_{t,n_{i,j}^{kk}(t)} \qquad (15)$$

$$\hat{\theta}_{i,j}^{kk^*}(t) \leq \theta_{i,j}^{kk}(t) + 2C_{t,n_{i,j}^{kk}(t)} \qquad (16)$$

We bound the probability of events (14) and (15) using Chernoff-Hoeffding bound),

$$\mathbb{P}\{\hat{\theta}_{i,j}^{kk^*}(t) \leq \theta_{i,j}^{kk^*}(t) - C_{t,n_{i,j}^{kk^*}(t)}\} \leq e^{-4} \qquad (17)$$

$$\mathbb{P}\{\hat{\theta}_{i,j}^{kk}(t) \leq \theta_{i,j}^{kk}(t) - C_{t,n_{i,j}^{kk}(t)}\} \leq e^{-4} \qquad (18)$$

For $l \geq \lceil \frac{8lnn}{(\Delta_{i,j}^{kk}(t))^2} \rceil$, (16) is false. In fact

$$\begin{aligned}
&\hat{\theta}_{i,j}^{kk^*}(t) - \theta_{i,j}^{kk}(t) - 2C_{t,n_{i,j}^{kk}(t)} \\
&= \hat{\theta}_{i,j}^{kk^*}(t) - 2\sqrt{\frac{2\ln t}{n_{i,j}^{kk}(t)}} \\
&\geq \hat{\theta}_{i,j}^{kk^*}(t) - \theta_{i,j}^{kk}(t) - 2\sqrt{\frac{2(\Delta_{i,j}^{kk}(t))^2 lnt}{8lnn}} \\
&= \hat{\theta}_{i,j}^{kk^*}(t) - \theta_{i,j}^{kk}(t) - \Delta_{i,j}^{kk}(t) = 0 \qquad (19)
\end{aligned}$$

If we let $l = \lceil \frac{8lnn}{(\Delta_{i,j}^{kk}(t))^2} \rceil$, then (16) is false for all $kk(t)$. Therefore,

$$\begin{aligned}
\mathbb{E}[T_{i,j}(n)] &\leq \frac{8lnn}{(\Delta_{min}^{i,j})^2} + M\sum_{t=1}^{\infty}\sum_{n_{i,j}^{kk^*}=1}^{t-1}\sum_{n_{i,j}^{kk(t)}=l}^{t-1} 2t^{-4} \\
&\leq \frac{8lnn}{(\Delta_{min}^{i,j})^2} + M(1+\frac{\pi^2}{3}) \qquad (20)
\end{aligned}$$

So under our scheme,

$$\begin{aligned}
\mathfrak{R}_1^{\pi}(\theta; n) &= n\sum_{(m,k)\in kk^*}\theta_{m,k} - \mathbb{E}^{\pi}[\sum_{t=1}^{n} S_{\pi(t)}(t)] \\
&= n\theta_{kk^*} - \mathbb{E}^{\pi}[\sum_{t=1}^{n} S_{\pi(t)}(t)] \\
&= \sum_{kk:\theta_{kk}<\theta_{kk^*}} \Delta_{kk}\mathbb{E}[T_{kk}(n)] \\
&\leq \Delta_{max}\sum_{kk:\theta_{kk}<\theta_{kk^*}} \mathbb{E}[T_{kk}(n)] \\
&= \Delta_{max}\sum_{i=1}^{M}\sum_{j=1}^{K}\mathbb{E}[T_{i,j}(n)] \\
&\leq [\sum_{i=1}^{M}\sum_{j=1}^{K}\frac{8lnn}{(\Delta_{min}^{i,j})^2} + M^2K(1+\frac{\pi^2}{3})]\Delta_{max} \\
&\leq [\frac{8MKlnn}{(\Delta_{min})^2} + M^2K(1+\frac{\pi^2}{3})]\Delta_{max} \qquad (21)
\end{aligned}$$

■

V. DSMU-*r*MAB: DISTRIBUTED STABLE MATCHING MULTI-USER ROBUST MAB FOR RELAY SELECTION

Temporal changes in reward distribution structure are intrinsic characteristics of problems in many application domains. Reward distribution structures under communication conditions uncertainty often involve trade-offs between learning about users' sensitivities to communication condition variations and earning short-term revenues. In this section, we focus on a MU-MAB formulation that allows for a broad range of temporal uncertainties in rewards due to the varying demands of the environment. A distributed stable matching multi-user robust MAB algorithm, DSMU-*r*MAB, is proposed to overcome abrupt changes in communication conditions under non-stationary MAB settings where the unknown reward distributions undergo changes in time and eliminates collisions among users through a one-to-one user-relay matching policy. In the presence of uncertainty, an agent that faces a sequence of decisions needs to judiciously use information collected from past observations when trying to optimize future actions. Knowing that undetected changes will lead to severe inaccuracies in estimation, the agent needs to discount the weight of older demand observations while estimating the demand curve in evolving environments.

The fundamental problem of multiple users contending for relay selection over multiple relays in UASNs has been formulated as a DSMU-*r*MAB problem. The goal is to design distributed online learning policies that incur minimal regret. The DSMU-*r*MAB problem we consider has the following key features. (a) For the purpose of considering dynamic environmental changes, the problem of relay selection can be modeled as non-stationary bandit problems where the distributions of rewards change abruptly at unknown time instants. Discounted-UCB(D-UCB), which was proposed in [37], is adequately successful when used to model evolving environments where the reward distributions undergo changes in time, but there have been no analyses of multi-user situations. We based the relay selection scheme on D-UCB and considered a set-up where there are multi-source nodes.

(b) To solve the competition among source node users of UASNs, we apply the stable matching theorem to allocate relays that effectively avoid multi-user collisions.

To estimate the instantaneous expected reward, the D-UCB scheme averages past rewards with a discount factor giving more weight to recent observations. In particular, D-UCB is a variant of the UCB policies that relies on a discount factor $\gamma \in (0, 1)$. This scheme constructs a index $b_{m,k}(t) = \bar{n}_{m,k}(t) + C_{m,k}(t)$ for the instantaneous expected reward, where the discounted exploration bonus is $C_{m,k}(t) = 2\sqrt{\xi ln n_t(\gamma)/\bar{n}_{m,k}(t)}$, with $n_t(\gamma) = \sum_{m=1}^{M}\sum_{k=1}^{K}\bar{n}_{m,k}(t)$, for an appropriate parameter ξ. At each time t, after an user-relay pair $(m, k) \in kk(t)$ is played, then $(\hat{\theta}'(m, k))_{M\times K}$ and $(\bar{n}_{m,k})_{M\times K}$ are update as follows: where the discounted empirical average and discounted number of times are given by

$$\begin{cases} \hat{\theta}'_{m,k}(t) = \dfrac{1}{\bar{n}_{m,k}(t)} X_{m,k}(s)\sum_{s=1}^{t}\gamma^{t-s}\mathbb{I}_{\{(m,k)\in kk(s)\}} \\ \bar{n}_{m,k}(t) = \sum_{s=1}^{t}\gamma^{t-s}\mathbb{I}_{\{(m,k)\in kk(s)\}} \end{cases} \tag{22}$$

Our proposed policy, which we refer to as a distributed stable matching multi-user robust MAB algorithm, is shown in Algorithm 3. We propose the subroutine presented in Algorithm 2 to solve the relevant distributed stable matching problem.

Algorithm 3 DSMU-rMAB

1: // INITIALIZATION
2: **for** $t = 1 \to K$ **do**
3: **for** $m = 1 \to M$ **do**
4: Select relay k such that $k = ((m + t) mod K) + 1$;
5: $\hat{\theta}'_{m,k}(t) = X_{m,k}(t)$;
6: $\bar{n}_{m,k}(t) = 1$;
7: **end for**
8: **end for**
9: // MAIN LOOP
10: **while** 1 **do**
11: $t = K + 1$ do
12: Run algorithm 2 to get a set contains M user-relay pairs that maximizes

$$\sum_{(p,q)\in kk(t)} \hat{\theta}'_{p,q}(t) + 2\sqrt{\xi ln n_t(\gamma)/\bar{n}_{p,q}(t)} \tag{23}$$

13: Update $(\hat{\theta}'_{m,k})_{M\times K}$, $(\bar{n}_{m,k})_{M\times K}$ accordingly.
14: **end while**

Now we provide the analysis of the upper-bound on the regret of DSMU-rMAB. Let Υ_n denote the number of breakpoints (we consider abruptly changing environments: the distributions of rewards remain constant during periods and change at unknown time slots) before time n. We denote by $\mathbb{E}_\gamma$ and $\mathbb{P}_\gamma$ the expectation and probability distribution under the our scheme DSMU-rMAB using the discount factor γ.

Theorem 2: The regret of DSMU-rMAB is

$$M^2K(1 + M\lceil n(1-\gamma)\rceil A(\gamma)\gamma^{-\frac{1}{1-\gamma}} + M\Upsilon_n D(\gamma) + 2M(\tau - K + \lceil \frac{ln\frac{1}{1-\gamma}}{ln(1+\eta)}\rceil \frac{n(1-\gamma)}{1-\gamma^{\frac{1}{1-\gamma}}})) \tag{24}$$

Proof: We introduce $T_{i,j}(n)$ as a counter after the initialization period. It is updated in the following way:

At each time slot after the initialization period, when non-optimal set $kk(t)$ is played at time t, there must be at least one user-relay pair $(i, j) \in kk(t)$, such that $(i, j) \notin kk^*(t)$. If there is only one such pair, $T_{i,j}(n)$ is increased by 1. If there are multiple such pairs, we arbitrarily pick one, say (p, q), and increment $T_{p,q}$ by 1.

Each time when a non-optimal set is picked, exactly one element in $(T_{i,j}(n))_{M\times K}$ is incremented by 1. This implies that the total number that we have played the non-optimal sets is equal to the summation of all counters in $(T_{i,j}(n))_{M\times K}$. Therefore, we have:

$$\sum_{kk:\theta_{kk}<\theta_{kk^*}} \mathbb{E}_\gamma[T_{kk}(n)] = \sum_{i=1}^{M}\sum_{j=1}^{K}\mathbb{E}_\gamma[T_{i,j}(n)] \tag{25}$$

The number of times a user-relay pair (i, j) that contain in a suboptimal arm is played is:

$$\begin{aligned} T_{i,j}(n) &= 1 + \sum_{t=K+1}^{n}\{\hat{\bar{\theta}}'_{kk^*(t-1)}(t-1) + C_{t-1,\bar{n}_{kk^*}(t-1)} \\ &\le \hat{\bar{\theta}}'_{kk(t-1)}(t-1) + C_{t-1,\bar{n}_{kk}(t-1)}, \bar{n}^{kk}_{i,j} < A(\gamma)\} \\ &\quad + \sum_{t=K+1}^{n}\{\hat{\bar{\theta}}'_{kk^*(t-1)}(t-1) + C_{t-1,\bar{n}_{kk^*}(t-1)} \\ &\le \hat{\bar{\theta}}'_{kk(t-1)}(t-1) + C_{t-1,\bar{n}_{kk}(t-1)}, \bar{n}^{kk}_{i,j} \ge A(\gamma)\} \end{aligned} \tag{26}$$

where $A(\gamma) = 16\xi ln n_n(\gamma)/(\Delta^{i,j}_{min})^2$. Note that $\hat{\bar{\theta}}'_{kk^*(t-1)}(t-1) + C_{t-1,\bar{n}_{kk^*}(t-1)} \le \hat{\bar{\theta}}'_{kk(t-1)}(t-1) + C_{t-1,\bar{n}_{kk}(t-1)}$, for $j\epsilon[1, K]$, at least one of the following must hold:

$$\begin{aligned} &\hat{\theta}'^{kk^*}_{1,j}(t) + C_{t,\bar{n}^{kk^*}_{1,j}(t)} \le \hat{\theta}'^{kk}_{1,j}(t) + C_{t,\bar{n}^{kk}_{1,j}(t)}, \\ &\hat{\theta}'^{kk^*}_{2,j}(t) + C_{t,\bar{n}^{kk^*}_{2,j}(t)} \le \hat{\theta}'^{kk}_{2,j}(t) + C_{t,\bar{n}^{kk}_{2,j}(t)}, \\ &\vdots \\ &\hat{\theta}'^{kk^*}_{M,j}(t) + C_{t,\bar{n}^{kk^*}_{M,j}(t)} \le \hat{\theta}'^{kk}_{M,j}(t) + C_{t,\bar{n}^{kk}_{M,j}(t)}. \end{aligned}$$

Formula (26) can be written:

$$\begin{aligned} T_{i,j}(n) &\le 1 + M\sum_{t=K+1}^{n}\{\hat{\theta}'^{kk^*}_{i,j}(t) + C_{t,\bar{n}^{kk^*}_{i,j}(t)} \\ &\le \hat{\theta}'^{kk}_{i,j}(t) + C_{t,\bar{n}^{kk}_{i,j}(t)}, \bar{n}^{kk}_{i,j} < A(\gamma)\} \\ &\quad + M\sum_{t=K+1}^{n}\{\hat{\theta}'^{kk^*}_{i,j}(t) + C_{t,\bar{n}^{kk^*}_{i,j}(t)} \\ &\le \hat{\theta}'^{kk}_{i,j}(t) + C_{t,\bar{n}^{kk}_{i,j}(t)}, \bar{n}^{kk}_{i,j} \ge A(\gamma)\} \end{aligned} \tag{27}$$

Theorem 3: [37] Let $i \in \{1,\dots,M\}$, $j \in \{1,\dots,K\}$; for any positive integer τ, when $\gamma = 1$, let $\bar{n}_{i,j}^{kk}(t-\tau:t) = \sum_{t-\tau+1}^{t} \mathbb{I}_{\{(i,j)\in kk(t)\}}$. Then for any positive a,

$$\sum_{t=K+1}^{n} \mathbb{I}_{\{(i,j)\in kk(t),\bar{n}_{i,j}^{kk}(t-\tau:t)<a\}} \leq \lceil n/\tau \rceil a. \tag{28}$$

Thus, for any $\tau \geq 1$, $\gamma \in (0,1)$ and $A > 0$, $\sum_{t=K+1}^{n} \mathbb{I}_{\{(i,j)\in kk(t),\bar{n}_{i,j}^{kk}(t)<A\}} \leq \lceil n/\tau \rceil A\gamma^{-\tau}$.

Using Theorem 3, we upper-bound the first sum in the RHS of (27) as $\sum_{t=K+1}^{n} \{\hat{\theta'}_{i,j}^{kk^*}(t) + C_{t,\bar{n}_{i,j}^{kk^*}(t)} \leq \hat{\theta'}_{i,j}^{kk}(t) + C_{t,\bar{n}_{i,j}^{kk}(t)}, \bar{n}_{kk_{i,j}}(t) < A(\gamma)\} \leq \lceil n(1-\gamma) \rceil A(\gamma)\gamma^{-\frac{1}{1-\gamma}}$.

For a number of rounds $D(\gamma)$ which depends on γ following a breakpoint, the estimates of the expected rewards can be poor for $D(\gamma) = ln((1-\gamma)\xi lnn_K(\gamma)/ln(\gamma)$ rounds, where $n_K(\gamma) = \sum_{i=1}^{M}\sum_{j=1}^{K} \bar{n}_{i,j}^{kk}(K)$. We denote by $\mathcal{T}(\gamma)$ the set of all indices $t \in \{K+1,\dots,n\}$, if it does not follow too soon after a state transition such that for all integers $s \in [t-D(\gamma), t]$, for all user-relay pairs (i,j), $\theta_{i,j}^{kk}(s) = \theta_{i,j}^{kk}(t)$. This leads to the following bound:

$$\begin{aligned}
&\sum_{t=K+1}^{n} \{\hat{\theta'}_{i,j}^{kk^*}(t) + C_{t,\bar{n}_{i,j}^{kk^*}(t)} \leq \hat{\theta'}_{i,j}^{kk}(t) + C_{t,\bar{n}_{i,j}^{kk}(t)}, \\
&\qquad \bar{n}_{i,j}^{kk}(t) \geq A(\gamma)\} \\
\leq\ & \Upsilon_n D(\gamma) + \sum_{t\in\mathcal{T}(\gamma)} \{\hat{\theta'}_{i,j}^{kk^*}(t) + C_{t,\bar{n}_{i,j}^{kk^*}(t)} \leq \hat{\theta'}_{i,j}^{kk}(t) \\
&+ C_{t,\bar{n}_{i,j}^{kk}(t)}, \bar{n}_{i,j}^{kk}(t) \geq A(\gamma)\}
\end{aligned} \tag{29}$$

Then we can obtain:

$$\begin{aligned}
T_{i,j}(n) \leq\ & 1 + M\lceil n(1-\gamma)\rceil A(\gamma)\gamma^{-\frac{1}{1-\gamma}} + M\Upsilon_n D(\gamma) \\
&+ M\sum_{t\in\mathcal{T}(\gamma)} \{\hat{\theta'}_{i,j}^{kk^*}(t) + C_{t,\bar{n}_{i,j}^{kk^*}(t)} \leq \hat{\theta'}_{i,j}^{kk}(t) \\
&+ C_{t,\bar{n}_{i,j}^{kk}(t)}, \bar{n}_{i,j}^{kk}(t) \geq A(\gamma)\}
\end{aligned} \tag{30}$$

Now observe that equation $\hat{\theta'}_{i,j}^{kk^*}(t) + C_{t,\bar{n}_{i,j}^{kk^*}(t)} \leq \hat{\theta'}_{i,j}^{kk}(t) + C_{t,\bar{n}_{i,j}^{kk}(t)}$ implies that at least one of the following must hold

$$\theta_{i,j}^{kk^*}(t) - \theta_{i,j}^{kk}(t) \leq 2C_{t,\bar{n}_{i,j}^{kk}(t)} \tag{31}$$

$$\hat{\theta'}_{i,j}^{kk}(t) \geq \theta_{i,j}^{kk}(t) + C_{t,\bar{n}_{i,j}^{kk}(t)} \tag{32}$$

$$\hat{\theta'}_{i,j}^{kk^*}(t) \leq \theta_{i,j}^{kk^*}(t) - C_{t,\bar{n}_{i,j}^{kk^*}(t)} \tag{33}$$

Note for the choice of $A(\gamma)$ given above, we have $C_{t,\bar{n}_{i,j}^{kk}(t)} \leq 2\sqrt{(\xi lnn_t(\gamma))/A(\gamma)} \leq \Delta_{min}^{i,j}/2$ which implies $\theta_{i,j'}^{kk^*}(t) - \theta_{i,j}^{kk}(t) - 2C_{t,\bar{n}_{i,j}^{kk}(t)} \geq 0$, then (31) is false.

Instead of using a Chernoff-Hoeffding bound, we bound the probability of events (32) and (33) using a novel tailored-made control on a self-normalized mean of the rewards with a random number of summands, which is stated in Theorem 4.

Theorem 4 [37]: For all integers t and all δ, $\eta > 0$,

$$\mathbb{P}\left(\frac{S_t(\gamma) - L_t(\gamma)}{\sqrt{N_t(\gamma^2)}} > \delta\right) \leq \lceil \frac{lnn_t(\gamma)}{ln(1+\eta)} \rceil exp(-2\delta^2(1-\frac{\eta^2}{16})). \tag{34}$$

We show that for $t \in \mathcal{T}(\gamma)$, that is at least $D(\lambda)$ rounds after a breakpoint, the expected rewards of all arms are well estimated with high probability. The idea is the following: we upper-bound the probability of (32) and (33) by separately considering the fluctuations of $\hat{\theta'}_{i,j}^{kk}(t)$ around $L_{i,j}^{kk}(t)/\bar{n}_{i,j}^{kk}(t)$, and the 'bias' $L_{i,j}^{kk}(t)/\bar{n}_{i,j}^{kk}(t) - \theta_{i,j}^{kk}(t)$, where $L_{i,j}^{kk}(t) = \sum_{s=1}^{t} \gamma^{t-s}\mathbb{I}_{\{(i,j)\in kk(t)\}}\theta_{i,j}^{kk}(s)$.

Note that $L_{i,j}^{kk}(t)/\bar{n}_{i,j}^{kk}(t)$, as a convex combination of elements $\theta_{i,j}^{kk}(s)$. Hence, $|L_{i,j}^{kk}(t)/\bar{n}_{i,j}^{kk}(t) - \theta_{i,j}^{kk}(s)| \leq 1$. For $t \in \mathcal{T}(\gamma)$,

$$\begin{aligned}
&|L_{i,j}^{kk}(t) - n_t(\gamma)\theta_{i,j}^{kk}(t)| \\
&= |\sum_{s=1}^{t-D(\gamma)} \gamma^{t-s}(\theta_{i,j}^{kk}(s) - \theta_{i,j}^{kk}(t))\mathbb{I}_{\{(i,j)\in kk(t)\}}| \\
&\leq \sum_{s=1}^{t-D(\gamma)} \gamma^{t-s}|\theta_{i,j}^{kk}(s) - \theta_{i,j}^{kk}(t)| \\
&\leq \gamma^{D(\gamma)}\bar{n}_{i,j}^{kk}(t - D(\gamma)).
\end{aligned} \tag{35}$$

Note that $\bar{n}_{i,j}^{kk}(t - D(\gamma)) \leq (1-\gamma)^{-1}$, we get that $|L_{i,j}^{kk}(t)/\bar{n}_{i,j}^{kk}(t) - \theta_{i,j}^{kk}(s)| \leq \gamma^{D(\gamma)}((1-\gamma)n_t(\gamma))^{-1}$. Altogether, $|\frac{L_{i,j}^{kk}(t)}{\bar{n}_{i,j}^{kk}(t)} - \theta_{i,j}^{kk}(t)| \leq (1 \wedge \frac{\gamma^{D(\gamma)}}{(1-\gamma)n_t(\gamma)})$. The elementary inequality $1 \wedge x \leq \sqrt{x}$. Hence, we obtain for $t \in \mathcal{T}(\gamma)$:

$$\begin{aligned}
|\frac{L_{i,j}^{kk}(t)}{\bar{n}_{i,j}^{kk}(t)} - \theta_{i,j}^{kk}(t)| &\leq \sqrt{\frac{\gamma^{D(\gamma)}}{(1-\gamma)n_t(\gamma)}} \\
&\leq \sqrt{\frac{\xi lnn_K(\gamma)}{\bar{n}_{i,j}^{kk}(t)}} \leq \frac{1}{2}C_{t,\bar{n}_{i,j}^{kk}(t)}.
\end{aligned} \tag{36}$$

Note that for $t \in \mathcal{T}(\gamma)$:

$$\begin{aligned}
&\mathbb{P}_\gamma(\hat{\theta'}_{i,j}^{kk}(t) \geq \theta_{i,j}^{kk}(t) + C_{t,\bar{n}_{i,j}^{kk}(t)}) \\
&\leq \mathbb{P}_\gamma(\hat{\theta'}_{i,j}^{kk}(t) \geq \theta_{i,j}^{kk}(t) + \sqrt{\frac{\xi lnn_K(\gamma)}{N_{i,j}^{kk}(t)}} \\
&\quad + |\frac{L_{i,j}^{kk}(t)}{\bar{n}_{i,j}^{kk}(t)} - \theta_{i,j}^{kk}(t)|) \\
&\leq \mathbb{P}_\gamma(\hat{\theta'}_{i,j}^{kk}(t) - \frac{L_{i,j}^{kk}(t)}{\bar{n}_{i,j}^{kk}(t)} > \sqrt{\frac{\xi lnn_K(\gamma)}{\bar{n}_{i,j}^{kk}(t)}}).
\end{aligned} \tag{37}$$

We denote by $S_{i,j}^{kk}(t)$ the discounted total reward obtained with user-relay pair (i,j). We bound the probability of events (37)

using Theorem 4 and the fact that $\bar{n}_{i,j}^{kk}(t) \geq \bar{n'}_{i,j}^{kk}(t)$, where $\bar{n'}_{i,j}^{kk}(t) = \sum_{s=1}^{t} \gamma^{2(t-s)} \mathbb{I}_{\{(i,j)\in kk(s)\}}$, we can get:

$$\begin{aligned}
&\mathbb{P}_\gamma(\hat{\theta'}_{i,j}^{kk}(t) \geq \theta_{i,j}^{kk}(t) + C_{t,\bar{n}_{i,j}^{kk}(t)}) \\
&\leq \mathbb{P}_\gamma\left(\frac{S_{i,j}^{kk}(t) - L_{i,j}^{kk}(t)}{\sqrt{\bar{n'}_{i,j}^{kk}(t)}} > \sqrt{\frac{\xi N_t(\gamma, kk_{i,j}) ln n_t(\gamma)}{\bar{n'}_{i,j}^{kk}(t)}}\right) \\
&\leq \mathbb{P}_\gamma\left(\frac{S_{i,j}^{kk}(t) - L_{i,j}^{kk}(t)}{\sqrt{\bar{n'}_{i,j}^{kk}(t)}} \geq \sqrt{\xi ln n_t(\gamma)}\right) \\
&\leq \lceil \frac{ln n_t(\gamma)}{ln(1+\eta)} \rceil n_t(\gamma)^{-2\xi(1-\frac{\eta^2}{16})} \qquad (38)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_\gamma[T_{i,j}(n)] \leq\ & 1 + M\lceil n(1-\gamma)\rceil A(\gamma)\gamma^{-1/(1-\gamma)} + MD(\gamma)\Upsilon_n \\
& + 2M \sum_{t\in\mathcal{T}(\gamma)} \lceil \frac{ln n_t(\gamma)}{ln(1+\eta)} \rceil n_t(\gamma)^{-2\xi(1-\frac{\eta^2}{16})}. \qquad (39)
\end{aligned}$$

When $\Upsilon_n \neq 0, \gamma < 1$. As $\xi > 0.5$, we take $\eta = 4\sqrt{1 - 1/2\xi}$, for that choice, with $\tau = \frac{1}{1-\gamma}$,

$$\begin{aligned}
&\sum_{t\in\mathcal{T}(\gamma)} \lceil \frac{ln n_t(\gamma)}{ln(1+\eta)} \rceil n_t(\gamma)^{-2\xi(1-\frac{\eta^2}{16})} \\
&\leq \tau - K + \sum_{t=\tau}^{n} \lceil \frac{ln n_\tau(\gamma)}{ln(1+\eta)} \rceil n_\tau(\gamma)^{-1} \\
&\leq \tau - K + \lceil \frac{ln n_\tau(\gamma)}{ln(1+\eta)} \rceil \frac{n}{n_\tau(\gamma)} \\
&\leq \tau - K + \lceil \frac{ln \frac{1}{1-\gamma}}{ln(1+\eta)} \rceil \frac{n(1-\gamma)}{1-\gamma^{\frac{1}{1-\gamma}}} \qquad (40)
\end{aligned}$$

So under our scheme,

$$\begin{aligned}
&\mathfrak{R}_2^\pi(\Theta(t); n) \\
&= \sum_{t=1}^{n} \sum_{(m,k)\in kk^*(t)} \theta_{m,k}(t) - \mathbb{E}^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)] \\
&= \sum_{t=1}^{n} \theta_{kk^*(t)} - \mathbb{E}^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)] \\
&\leq M \sum_{kk:\theta_{kk}<\theta_{kk*}} \mathbb{E}_\gamma[T_{kk}(n)] \\
&= M \sum_{i=1}^{M}\sum_{j=1}^{K} \mathbb{E}_\gamma[T_{i,j}(n)] \\
&\leq M^2K(1 + M\lceil n(1-\gamma)\rceil A(\gamma)\gamma^{-\frac{1}{1-\gamma}} + M\Upsilon_n D(\gamma) \\
&\quad + 2M(\tau - K + \lceil \frac{ln \frac{1}{1-\gamma}}{ln(1+\eta)} \rceil \frac{n(1-\gamma)}{1-\gamma^{\frac{1}{1-\gamma}}})) \qquad (41)
\end{aligned}$$

■

VI. PERFORMANCE EVALUATION

A. OVERHEAD ANALYSIS OF RELAY SELECTION IN MU-MAB FRAMEWORK

In the MU-MAB framework, we need the average of all the observed rewards of relay k by user m and number of times that relay k is selected by user m to calculate the UCB index for each user-relay pair (m, k). Although each node has to carry out frequent computations for the UCB index, like other machine learning algorithms [38], the computations of the decision-making parameters are simple and their delays and power consumptions are much smaller than for acoustic communications. Hence, the computational overhead of the UCB index is ignored.

In centralized MU-MAB algorithms [23], [30], the tuple of $\langle$`index_m, index_k, Value_`$b_{m,k}(t)\rangle$ of m is referred to as a `QUEST` message, where `index_`m represents the sender's (i.e., source node) *ID*, `index_`k represents the relay's *ID*, and `Value_`$b_{m,k}(t)$ represents the UCB index at t of the user-relay pair. At each time t, each source node user sends its K `QUEST` messages to the coordinator. The coordinator can then be in charge of announcing the non-conflicting relays to be used by each user for each decision period after running the Hungarian allocation algorithm. Finally, the coordinator needs to broadcast an `ALLOCATION` message with format $\langle$`index_m, index_k`$\rangle$. Hence, the overhead of the centralized MU-MAB algorithm arises from two places: the `QUEST` message and the `ALLOCATION` message. The message complexity of centralized MU-MAB algorithms is $O(M \times K + M)$. In contrast to these centralized MU-MAB algorithms, our DSMU-MAB and DSMU-rMAB algorithms are fully distributed. It is obvious that the nodes are completely dependent on local information to make decisions and do not need the entirety of network topology information. Obviously, in the implementation of the algorithm, each node only needs to broadcast a `MATCHING` message with format $\langle$`index_m, index_k`$\rangle$. The message complexity of DSMU-MAB is $O(M)$, which is the same as that for DSMU-rMAB.

B. SIMULATION AND ANALYSIS OF DSMU-MAB

1) METHODOLNY AND SIMULATION SETUP

In this experiment, we consider source nodes willing to exploit relay nodes with unknown expected reward patterns $\Theta = (\theta_{m,k})_{M\times K}$ and evaluate the nodes' transmission performances depending on network throughput, and then denote $\theta_{m,k}$ as the expected reward of relay k observed by user m. To simplify the process, we assume that follows a Gaussian distribution with parameter $\theta_{m,k}$.[4] We verified the feasibility of that distribution using *Watermark version* 1.0 dataset [39], it contains mass of time-varying acoustics channels data, which were measured in Norwegian waters in the frequency band 10 to 18 kHz.

[4]This Gaussian distribution assumption is only used for performance evaluation, and there is no prior reward distribution knowledge provided to DSMU-MAB during relay selection.

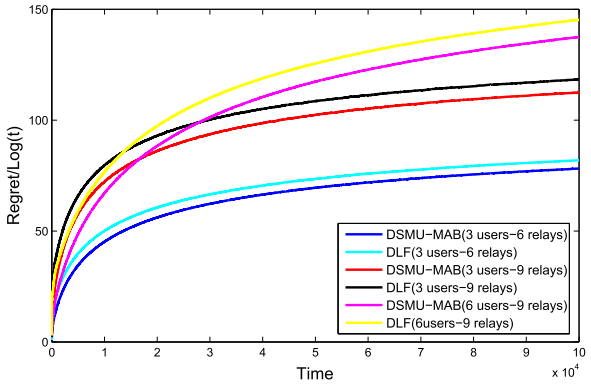


FIGURE 4. Normalized regret $\frac{\Re(n)}{ln(n)}$ vs. n time slots.

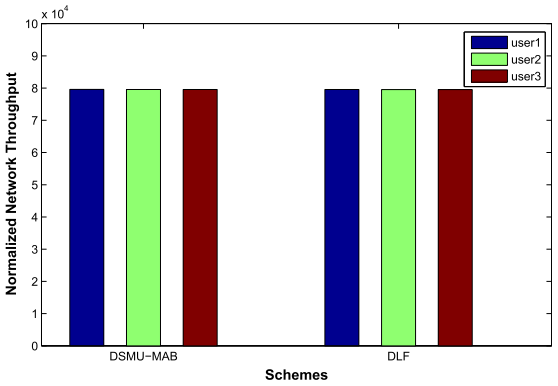


FIGURE 5. Different user networks' throughput.

TABLE 2. Performance comparison of several algorithms ($M = 3, S = 2, G = 1, K = 9$).

	type	ratio to optimal ratio
Optimal allocation scheme	centralized	100%
Greedy allocation scheme	distributed	37.20%
Random scheme	distributed	45.57%
Hungarian MU-MAB	centralized	99.44%
DSMU-MAB	distributed	99.41%
MLPS	centralized	97.60%

2) COMPARISONS OF DSMU-MAB AND DLF IN A SYMMETRICAL CASE

In this section, we first compare the performance of DSMU-MAB with the distributed MU-MAB algorithm DLF, which was only designed for symmetrical cases. In Fig. 4, we compare the normalized regrets $\frac{\Re(n)}{ln(n)}$ of our scheme and the DLF access scheme, which were determined by varying the number of users and relays to verify the performance of DSMU-MAB detailed earlier. We show the simulations of three sets of data, (i) we have $M = 3$ users, $K = 6$ relays, with $\Theta_{m,1:K} = (0.9, 0.8, 0.7, 0.6, 0.5, 0.4), m = 1, 2, 3$; (ii) we have $M = 3$ users, $K = 9$ relays, with $\Theta_{m,1:K} = (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1), m = 1, 2, 3$; (iii) we have $M = 6$ users, $K = 9$ relays, with $\Theta_{m,1:K} = (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1), m = 1, 2, 3, \ldots, 6$. It can be easily determined that the regret is uniformly logarithmic with time slot. As expected, our scheme can yield the least regret, because it can completely eliminate conflicts.

The advantage of DLF is that it enables fairness access for all users with same prioritizations in a symmetrical case. Thus, we also present a comparison of fairness. As shown in Fig. 5, we provide a bar chart for each user's cumulative normalized network throughput running on DSMU-MAB and DLF. We found that fairness among different users was reflected in both the DSMU-MAB and DLF schemes. Although in DSMU-MAB, we need not change rank of priority access for each user like DLF, the users could also compete fairly well for the best relay because of the inherent fairness of our scheme.

3) PERFORMANCE EVALUATION OF DSMU-MAB IN AN ASYMMETRICAL CASE

As mentioned before, asymmetrical cases are more common in multi-user UASNs relay selection. We therefore pay more attention to performance evaluation of DSMU-MAB in an asymmetrical case. In fact, in asymmetrical cases, some users also experience the same reward process on a given relay, and we call them symmetrical users. For simplicity, in this section, the number of symmetrical users is S, the number of groups of symmetrical users is G. Obviously, when $M = S$, $G = 1$, and the case is symmetrical, and when $S = 0$, $G = 0$, and the case is completely asymmetrical. Unlike other distributed MU-MAB algorithms, our scheme remains effective for asymmetric cases.

We compared DSMU-MAB with four different relay selection schemes for the case of $M = 3, S = 2, G = 1, K = 9$, namely, (i) a random scheme, wherein a relay is randomly selected from the K available at each slot with equal probability, (ii) a greedy allocation scheme in which users learn unknown channel parameters but select the relay providing the highest UCB index of selfishness, (iii) the MLPS scheme presented in [23] and (iv) the Hungarian centralized MU-MAB scheme presented in [30]. We consider a reward matrix Θ defined as:

$$\Theta = \begin{bmatrix} 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \\ 0.7 & 0.6 & 0.5 & 0.8 & 0.3 & 0.2 & 0.9 & 0.1 & 0.4 \end{bmatrix}.$$

As shown in Fig. 6, we provide a bar chart regarding the cumulative normalized network throughputs for the five schemes. The performance of our distributed scheme was close to that of the Hungarian centralized MU-MAB, better than that of MLPS, and had significant performance advantages over the random and greedy allocation schemes. In our configuration, we also define an optimal scheme as representing the ideal case, wherein instantaneous full CSI corresponding to all the user-relay pairs is available for matching the optimal user-relay pairs. In this scheme, it is assumed that the mean rewards for all the user-relay pairs are known a priori, and a genie always selects the optimal relay. The value of the optimal scheme can be used as the upper bound of the total cumulative normalized network throughput of each user.

TABLE 3. Hungarian MU-MAB($M = 3, S = 2, G = 1, K = 9$).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$m=1$	49897	47548	1545	468	222	130	85	60	45
$m=2$	49809	47638	1542	468	224	129	85	60	45
$m=3$	232	231	133	1731	62	46	97443	35	87

TABLE 4. Greedy allocation($M = 3, S = 2, G = 1, K = 9$).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$m=1$	97183	1724	496	234	135	87	61	45	35
$m=2$	97183	1728	497	231	133	87	61	45	35
$m=3$	497	231	134	1743	62	45	97166	35	87

TABLE 5. MLPS($M = 3, S = 2, G = 1, K = 9$).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$m=1$	93012	1123	4662	1	1	1	1	1	1192
$m=2$	6975	92862	151	1	1	1	1	1	1
$m=3$	1	1	1853	13898	1	1	78304	1	5934

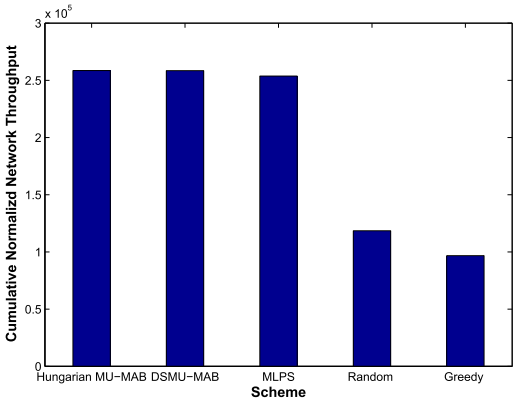


FIGURE 6. Total network throughput under five schemes ($M = 3, S = 2, G = 1, K = 9$).

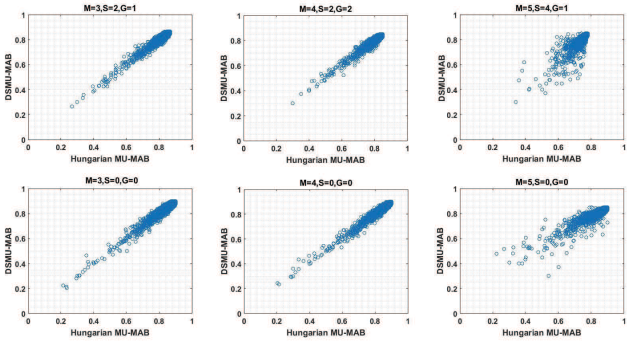


FIGURE 7. DSMU-MAB vs. Hungarian MU-MAB algorithm performance.

Table II lists the performance ratios of the different schemes to that of the optimal scheme based on experimental simulation. We showed that the rate achieved by DSMU-MAB was approximately 99.41% of the optimal rate. The difference between the Hungarian MU-MAB rate and our rate was at most 0.03%. In addition, the total cumulative normalized network throughput increased 54%over that of the random scheme.

The numbers of times that each relay was selected by each source node are shown in table III-table VII, and they clearly reflect that DSMU-MAB can significantly increase the best relay utilization rate. As expected, the greedy allocation scheme produced the worst performance because two symmetric users can select the best relay at the same time, as shown in table IV. That scheme cannot avoid communication conflicts among multiple symmetric users. In contrast to the greedy allocation scheme, our scheme confirms its ability to reduce the impact of collisions. The data for the random allocation scheme, whereby source node users select an arbitrary relay with equal probability, are shown in table VII and clearly reflect that the scheme can significantly increase relay utilization rate with low throughput.

In the above setting, our DSMU-MAB algorithm showed good performance that was similar to that of the Hungarian MU-MAB. To validate the applicability of DSMU-MAB, we also conducted simulations for other scenarios. First, we evaluated the cumulative regrets of DSMU-MAB and Hungarian MU-MAB and defined $d = \frac{regret\{\text{DSMU-MAB}\}}{regret\{\text{Hungarian MU-MAB}\}}$ to measure the performance gap between the two algorithms, as shown in table VIII. In six representative cases, our DSMU-MAB distributed algorithm had cumulative regrets similar to those of Hungarian MU-MAB. We then tested the average reward per user at each slot t, and scatter plots of the rewards for six cases are shown in Fig. 7. For most of the total $n = 10^5$ slots, the users had a similar average rewards at each slot for DSMU-MAB and Hungarian MU-MAB.

The simulation results above demonstrate that our DSMU-MAB distributed algorithm can achieve good performance in symmetrical and asymmetrical cases. For the symmetrical case, DSMU-MAB had a lower regret than that of existing distributed algorithm DLF and achieved fairness access on a par with DLF. Unlike DLF, DSMU-MAB can also be applied to asymmetrical cases and achieved performances in many cases similar to the centralized Hungarian MU-MAB in all above tests with a low level of overhead. In particular, although no prior CSI knowledge was provided to DSMU-MAB during relay selection, the algorithm achieved above 99% of the total rate of the optimal allocation, in which

TABLE 6. DSMU-MAB($M = 3, S = 2, G = 1, K = 9$).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$m=1$	50007	47440	1538	473	223	130	85	60	44
$m=2$	49460	47972	1558	468	224	128	85	60	45
$m=3$	471	232	133	1719	61	46	97216	35	87

TABLE 7. Random scheme($M = 3, S = 2, G = 1, K = 9$).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$m=1$	11013	11307	11064	11044	11167	11234	11169	10956	11046
$m=2$	11031	11290	10999	11164	10948	11069	11198	11010	11291
$m=3$	11153	11135	11171	11213	10972	11017	11105	11155	11079

TABLE 8. Regret comparison between DSMU-MAB and Hungarian MU-MAB($K = 9, n = 10^5$).

	$M=3, S=2, G=1$	$M=3, S=0, G=0$	$M=4, S=2, G=2$	$M=4, S=0, G=0$	$M=5, S=4, G=1$	$M=5, S=0, G=0$
d	1.0485	0.9996	1.0003	0.9997	1.1379	1.0013

instantaneous and full CSI corresponding to all the user-relay pairs was available.

C. SIMULATION AND ANALYSIS OF DSMU-rMAB IN CHANGING ENVIRONMENTS

We present simulation results for our proposed the DSMU-rMAB in changing environments. We assumed that the reward of relay k observed by user m at slot t also follows a Gaussian distribution, but the parameter $(\theta_{m,k})_{M\times K}$ suffers a change during the relay selection. We mainly considered "bursty" and "smooth" changes in $(\theta_{m,k})_{M\times K}$ to simulate the high dynamic changes in a shallow seabed or transient acoustic channel access from other artificial acoustic or natural acoustic systems.

In the first example, we consider "bursty" changes to simulate dynamic changes caused by wind speed changes and ship noise in a shallow seabed or by the sudden arrival of moving objects such as marine mammals and fish schools. We have $M = 3$ users, $K = 5$ relays, and $S = 3, G = 0$, the time horizon is set to $n = 10^4$. We consider the two breakpoint($t = 2000$ and $t = 5000$), for $t < 2000$, the expected rewards distribution matrix Θ is defined as:

$$\Theta = \begin{bmatrix} 0.9 & 0.8 & 0.7 & 0.6 & 0.5 \\ 0.6 & 0.5 & 0.8 & 0.9 & 0.7 \\ 0.5 & 0.7 & 0.9 & 0.8 & 0.6 \end{bmatrix}, \quad \text{for } 2000 \le t < 5000,$$

$$\Theta = \begin{bmatrix} 0.5 & 0.7 & 0.8 & 0.9 & 0.6 \\ 0.9 & 0.8 & 0.7 & 0.5 & 0.6 \\ 0.7 & 0.5 & 0.6 & 0.8 & 0.9 \end{bmatrix}, \quad \text{for } t \ge 5000,$$

$$\Theta = \begin{bmatrix} 0.8 & 0.9 & 0.5 & 0.7 & 0.6 \\ 0.5 & 0.6 & 0.8 & 0.7 & 0.9 \\ 0.7 & 0.8 & 0.9 & 0.6 & 0.5 \end{bmatrix}.$$

We represent the evolution of cumulated regret in Fig. 8. As seen in Fig. 8, DSMU-MAB performed with lower regret than DSMU-rMAB before $t = 2000$, however, the regret for DSMU-MAB converged with difficulty after the first breakpoint until the end of that trial. In contrast, the regret for DSMU-rMAB converged to a stable level although suffering two breakpoints, indicating that DSMU-rMAB can quickly concentrate their pulls on the optimal combination and is robust to bursty changes.

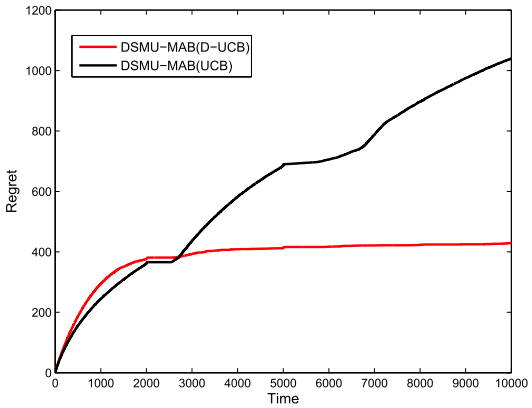


FIGURE 8. Normalized regret $\Re(n)$ vs. n time slots.

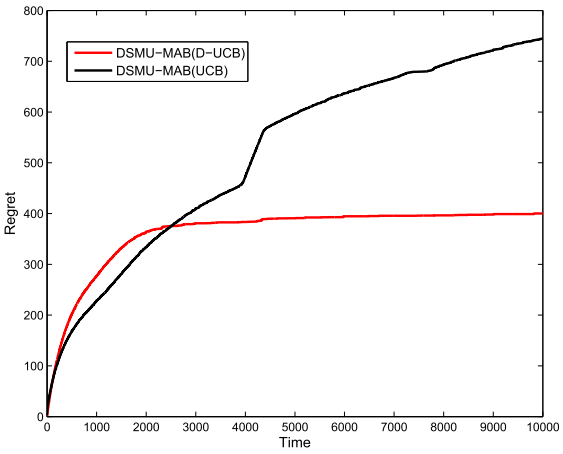


FIGURE 9. Normalized regret $\Re(n)$ vs. n time slots.

In the second example, we considered a "smooth" changes to simulate dynamic changes in a communication environment caused by spectral occupancy by other artificial acoustic systems. We tested the behaviors of DSMU-rMAB and DSMU-MAB by investigating their performances in a smoothly varying environment. That environment was made of $K = 5$ arms, and it was assumed that some elements of the reward matrix θ suffered a time-varying sinusoidal disturbance $\sigma(t) = sin(6\pi t/n)$. The best arm changed cyclically, and the transitions were smooth.

We consider an expected reward matrix θ defined as: $\Theta(t) = \begin{bmatrix} 0.6 & 0.7+0.2\sigma(t) & 0.5+0.4\sigma(t) & 0.4 & 0.9 \\ 0.8+0.1\sigma(t) & 0.5+0.1\sigma(t) & 0.3 & 0.9 & 0.6 \\ 0.6+0.2\sigma(t) & 0.6 & 0.9 & 0.5 & 0.3+\sigma(t) \end{bmatrix}$.

The evolutions of the cumulative regrets under the two policies are shown in Fig. 9; in this continuously evolving environment, the DSMU-MAB algorithms accumulated larger regrets, the convergences of which were difficult. However, DSMU-rMAB was robust to the continuous time-varying disturbance, and the regret converged at a stable level.

These modest and yet representative examples suggest that DSMU-*r*MAB can be successfully adapted to cope with changing environments and achieve better performance than DSMU-MAB, whether bursty or smoothly changing.

VII. CONCLUSIONS

In this paper, we studied the problem of multi-user relay allocations for UASNs and proposed for the first time a MU-MAB algorithm to efficiently solve the relay selection problem in a distributed manner without any knowledge regarding the nature of communication environments. Through theoretical analysis, we established that DSMU-MAB can achieve simple, distributed and efficient solutions without collisions among users and reduces the mass of information exchanged among users. At the same time, DSMU-*r*MAB, a derivative of DSMU-MAB, was proposed that can be robust to substantial changes in underwater communication environments. Numerical analysis showed that DSMU-MAB can be applied with decent performance to symmetrical and asymmetrical cases that can closely approach the optimal solution, and, moreover, DSMU-*r*MAB can be successfully robust to abrupt changes in environment.

REFERENCES

[1] G. Han, J. Jiang, N. Sun, and L. Shu, "Secure communication for underwater acoustic sensor networks," *IEEE Commun. Mag.*, vol. 53, no. 8, pp. 54–60, Aug. 2015.

[2] T. Yang, "Distributed underwater sensing: A paradigm change for the future," in *Advanced Materials*. New York, NY, USA: Springer, 2014, pp. 261–275.

[3] Y. Su, Y. Zhu, H. Mo, J.-H. Cui, and Z. Jin, "A joint power control and rate adaptation MAC protocol for underwater sensor networks," *Ad Hoc Netw.*, vol. 26, pp. 36–49, Mar. 2015.

[4] S. Al-Dharrab, M. Uysal, and T. M. Duman, "Cooperative underwater acoustic communications [Accepted From Open Call]," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 146–153, Jul. 2013.

[5] C. Murphy, J. M. Walls, T. Schneider, R. M. Eustice, M. Stojanovic, and H. Singh, "CAPTURE: A communications architecture for progressive transmission via underwater relays with eavesdropping," *IEEE J. Ocean. Eng.*, vol. 39, no. 1, pp. 120–130, Jan. 2014.

[6] J. Gomes *et al.*, "An overview of project compound: Cooperative communications and positioning in mobile underwater networks," in *Proc. Future Netw. Mobile Summit (FutureNetw)*, Jul. 2012, pp. 1–9.

[7] A. Doosti-Aref and A. Ebrahimzadeh, "Adaptive relay selection and power allocation for ofdm cooperative underwater acoustic systems," *IEEE Trans. Mobile Comput.*, vol. 17, no. 11, pp. 1–15, Jan. 2018.

[8] Y. Wei and D.-S. Kim, "Exploiting cooperative relay for reliable communications in underwater acoustic sensor networks," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, Oct. 2014, pp. 518–524.

[9] Y. Luo, L. Pu, Z. Peng, Z. Zhou, J.-H. Cui, and Z. Zhang, "Effective relay selection for underwater cooperative acoustic networks," in *Proc. IEEE 10th Int. Conf. Mobile Ad-Hoc Sensor Syst. (MASS)*, Oct. 2013, pp. 104–112.

[10] R. Gonen and E. Pavlov, "An incentive-compatible multi-armed bandit mechanism," in *Proc. 26th Annu. ACM Symp. Principles Distrib. Comput.*, 2007, pp. 362–363.

[11] D. He, W. Chen, L. Wang, and T.-Y. Liu, "Online learning for auction mechanism in bandit setting," *Decision Support Syst.*, vol. 56, pp. 379–386, Dec. 2013.

[12] W. Ding, T. Qiny, X.-D. Zhang, and T.-Y. Liu, "Multi-armed bandit with budget constraint and variable costs," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 232–238.

[13] G. Pini, A. Brutschy, G. Francesca, M. Dorigo, and M. Birattari, "Multi-armed bandit formulation of the task partitioning problem in swarm robotics," in *Proc. Int. Conf. Swarm Intell.*, 2012, pp. 109–120.

[14] P. Matikainen, P. M. Furlong, R. Sukthankar, and M. Hebert, "Multi-armed recommendation bandits for selecting state machine policies for robotic systems," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2013, pp. 4545–4551.

[15] P. Zhou and T. Jiang, "Toward optimal adaptive wireless communications in unknown environments," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3655–3667, May 2016.

[16] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5G small cells," *IEEE Wireless Commun.*, vol. 23, no. 3, pp. 64–73, Jun. 2016.

[17] N. Gulati and K. R. Dandekar, "Learning state selection for reconfigurable antennas: A multi-armed bandit approach," *IEEE Trans. Antennas Propag.*, vol. 62, no. 3, pp. 1027–1038, Mar. 2014.

[18] S. Abdulhadi, M. Jaseemuddin, and A. Anpalagan, "A survey of distributed relay selection schemes in cooperative wireless ad hoc networks," *Wireless Pers. Commun.*, vol. 63, no. 4, pp. 917–935, 2012.

[19] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, no. 1, pp. 9–15, Jan. 1962.

[20] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6184–6193, Dec. 2014.

[21] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: Indexablity and index policy," in *Proc. 5th IEEE Annu. Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw. Workshops (SECON)*, Jun. 2008, pp. 1–5.

[22] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.

[23] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. IEEE Symp. New Frontiers Dyn. Spect.*, Apr. 2010, pp. 1–9.

[24] S. Maghsudi and S. Stanczak, "Joint channel selection and power control in infrastructureless wireless networks: A multiplayer multiarmed bandit framework," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4565–4578, Oct. 2015.

[25] S. Maghsudi and S. Stańczak, "Joint channel allocation and power control for underlay D2D transmission," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2091–2096.

[26] B. Nikfar and A. J. H. Vinck, "Relay selection in cooperative power line communication: A multi-armed bandit approach," *J. Commun. Netw.*, vol. 19, no. 1, pp. 1–9, 2017.

[27] S. Shankar and M. Chitre, "Tuning an underwater communication link," in *Proc. MTS/IEEE OCEANS-Bergen*, Jun. 2013, pp. 1–9.

[28] D. M. Jayasuriya, "Adapting underwater physical link parameters using data driven algorithms," Ph.D. dissertation, 2010.

[29] X. Li, J. Liu, L. Yan, S. Han, and X. Guan, "Relay selection in underwater acoustic cooperative networks: A contextual bandit approach," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 382–385, Feb. 2017.

[30] Y. Gai, B. Krishnamachari, and M. Liu, "On the combinatorial multi-armed bandit problem with Markovian rewards," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2011, pp. 1–6.

[31] L. Huang, G. Zhu, X. Du, and K. Bian, "Stable multiuser channel allocations in opportunistic spectrum access," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 1715–1720.

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

[33] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Minimax policies for combinatorial prediction games," in *Proc. 24th Annu. Conf. Learn. Theory*, vol. 19, 2011, pp. 107–132.

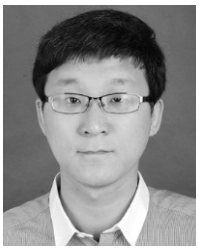
[34] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.
[35] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.
[36] A. Leshem, E. Zehavi, and Y. Yaffe, "Multichannel opportunistic carrier sensing for stable channel access control in cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 82–95, Jan. 2012.
[37] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2011, pp. 174–188.
[38] T. Hu and Y. Fei, "QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 6, pp. 796–809, Jun. 2010.
[39] P. van Walree, R. Otnes, and T. Jenserud, "The watermark manual and user's guide," Forsvarets Forskningsinstitutt, Tech. Rep., 2016.

XINBIN LI received the M.Sc. degree in control theory and control engineering from Yanshan University, China, in 1999, and the Ph.D. degree in general and fundamental mechanics from Peking University, China, in 2004. He is currently a Professor with the Institute of Electrical Engineering, Yanshan University, China. His research interests include nonlinear system and underwater acoustic networks.

JIAJIA LIU received the B.Eng. degree in electrical engineering and its automation from Yanshan University, Qinhuangdao, China, in 2015. She is currently pursuing the M.Sc. degree with the Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University. Her research interests include underwater acoustic relay networks and multi-armed bandit theory.

LEI YAN received the B.Eng. degree in electrical engineering and its automation from Yanshan University, Qinhuangdao, China, in 2013. He is currently pursuing the Ph.D. degree with the Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University.

His research interests include multi-armed bandit theory and design of algorithms for underwater acoustic sensor networks.



SONG HAN received the B.Eng. degree in automation from Xingtai College, Xingtai, China, in 2012. He is currently pursuing the Ph.D. degree with the Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao, China.

His research interests include game theory and resource allocation algorithm in underwater acoustic networks.

XINPING GUAN (F'18) received the Ph.D. degree in control and systems from the Harbin Institute of Technology, China. In 2007, he joined the Department of Automation, Shanghai Jiao Tong University, China, where he is currently the Distinguished University Professor and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China. He was the Leader of the prestigious Innovative Research Team, NSFC, in 2012. As a Principle Investigator, he has finished/been involved in many national key projects. He is also a Cyber Principle Investigator with the Cyber Joint Innovation Center founded by Zhejiang University, Tsinghua University, and Shanghai Jiao Tong University. He was appointed as the Changjiang Scholar by the Ministry of Education of China and the State-level Scholar of New Century Bai Qianwan Talent Program of China. He has authored and/or co-authored two research monographs, over 120 SCI indexed papers in the IEEE Transactions and other well-known international journals and numerous conference papers. His current research interests include wireless sensor networks, ground-air communication of aircrafts, and cognitive radio networks and their applications in industry. He is currently a Committee Member of the Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He was a recipient of the First Prize at the University Natural Science Award from the Ministry of Education of China in 2003 and 2007, respectively. He received the IEEE TRANSACTIONS ON FUZZY SYSTEMS Outstanding Paper Award in 2008 and the Second Prize at the National Natural Science Award of China. He received the National Outstanding Youth Award from the National Natural Science Foundation of China. He serves as an Associate Editor for the IEEE TRANSACTIONS ON SYSTEM, MAN AND CYBERNETICS-C, as an Editorial Board Committee Member for the several Chinese journals, and as an International Technical Committee Member for a lot of conferences.