PLOS ONE

RESEARCH ARTICLE

# Decentralized Opportunistic Spectrum Resources Access Model and Algorithm toward Cooperative Ad-Hoc Networks

**Ming Liu, Yang Xu*, Abdul-Wahid Mohammed**

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, P.R. China

* xuyang@uestc.edu.cn





**Citation:** Liu M, Xu Y, Mohammed A-W (2016) Decentralized Opportunistic Spectrum Resources Access Model and Algorithm toward Cooperative Ad-Hoc Networks. PLoS ONE 11(1): e0145526. doi:10.1371/journal.pone.0145526

**Editor:** Catalin Buiu, Politehnica University of Bucharest, ROMANIA

**Received:** September 11, 2015

**Accepted:** December 4, 2015

**Published:** January 4, 2016



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This research was sponsored by the National Natural Science Foundation of China (http://www.nsfc.gov.cn/) grants 61370151 and 61202211 to YX, the National Science and Technology Major Project of China (http://www.nmp.gov.cn/) grant 2015ZX03003012 to YX, the Central University Basic Research Funds Foundation of China grant ZYGX2014J055 to YX, and the Science and Technology on Electronic Information Control Laboratory Project. The funders had no role in study

## Abstract

Limited communication resources have gradually become a critical factor toward efficiency of decentralized large scale multi-agent coordination when both system scales up and tasks become more complex. In current researches, due to the agent's limited communication and observational capability, an agent in a decentralized setting can only choose a part of channels to access, but cannot perceive or share global information. Each agent's cooperative decision is based on the partial observation of the system state, and as such, uncertainty in the communication network is unavoidable. In this situation, it is a major challenge working out cooperative decision-making under uncertainty with only a partial observation of the environment. In this paper, we propose a decentralized approach that allows agents cooperatively search and independently choose channels. The key to our design is to build an up-to-date observation for each agent's view so that a local decision model is achievable in a large scale team coordination. We simplify the Dec-POMDP model problem, and each agent can jointly work out its communication policy in order to improve its local decision utilities for the choice of communication resources. Finally, we discuss an implicate resource competition game, and show that, there exists an approximate resources access tradeoff balance between agents. Based on this discovery, the tradeoff between real-time decision-making and the efficiency of cooperation using these channels can be well improved.

## Introduction

Communication resources always play a latent role in networked large-scale agent team coordination applications, such as multi-robots system, mobile sensor system, etc. With the expansion of the system, communication resources exert a momentous impact on the cooperative efficiency [1], and numerous attention from both industry and academia has been devoted to this research [2]. For instance, the utmost transfer rate of IEEE 802.11b protocol is 11Mbit/s, and with the insecurity of latency and packet loss, this may fail to meet the capacity requirement of large-scale robots carrying video equipment for surveillance in an open environment [3]. In our previous work [4], we found that, with the expansion of the team size, robots will

design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

compete the limited spectrum resources, which is a phenomenon also supported by other studies [5, 6]. On the other hand, different from Cognitive Radio (CR) [7], Mobile Sensor [8] and other traditional wireless communication researches, multi-agent system usually consist of multiple inexpensive agents, and without a strong central processing unit or resources pre-authorization, but with more incomplete channels observation and changing dynamics. In addition, there are no typical technical characteristics, such as a fixed base station or a central node to manage and distribute channels, etc. The major communication mode for most decentralized multi-agent system is Ad-hoc network [2]. However, the typical pre-authorization and consultative allocation approach cannot be applied in the dynamic tasks and agents' migration. In consequence, new concepts and strategies should be developed, and this is the main motivation proposed here.

As a main technical part of our research. In this paper, we model the decentralized multi-agent multi-channel access problem as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) problem. We use a continuous time Markov model to simulate the usage of channels while the constant slotted opportunity is used to support agents' interaction. In addition, we use a sample-based Partially Observable Markov Decision Process (POMDP) to simplify the model. Finally, based on game theory, we model and analyze implicit resource competition between agents, and prove the existence of equilibrium in an ideal state.

## State of the Art

Even though centralized channel resource allocation methods can provide some sort of optimal solutions, they are less effective in situations where the central point fails. For instance, typical auction-based algorithms generally have low communication requirements [9], and the negotiation process, in addition, can degrade in overall efficiency as communication deteriorates [10]. It has been shown in [11] that spatial channels opportunity allocation is equivalent to a graph coloring problem, which objective is to obtain colors assignment that maximizes the utility. But obtaining the optimal coloring is generally known to be NP-hard.

Opportunistic Spectrum Access (OSA) [12] and Opportunistic Spectrum Sharing (OSS) [13] are widely adopted in most recent researches, and several investigations have modeled OSA problems as a POMDP model [14]. Basic OSA concept is described as an agent, which can identify and access idle frequency bands and obtain maximized rewards. Many decentralized methods have referenced the design of POMDP, varying reliance on schemes and can only handle intermittent communication resource scheduling. Reinforcement learning (RL) [15] is a paradigm to solve POMDP problems, and it is inspired by a learning theory which has good performance in multi-robots decision applications [16, 17]. For most RL-based multi-agent systems, the rewards are more achieved by long-team learning, which is the expected accumulated reward that the agent expects to receive in the future under the policy, and can be specified by update value function. However, for the fixed utility function design, time restrains, interaction and observation limited applications, RL is restricted.

Game Theory provides another approach to OSA. Stochastic game [18] as an extension of Game Theory, can improve the capability to solve the OSA problems, and a deeper analysis between the game and the graph-based method is noted in [19]. It is important to note that in many situations, states of the system cannot be observed completely. Therefore, some researches adopt the definition of Partially Observable Stochastic Game (POSG), and a cooperative case of POSG, namely Dec-POMDP [20]. Although some efforts have been made in building heuristic algorithms to solve this intrinsic NEXP-complete problem [21], it is still less feasible obtaining optimal results in a limited time with the partial observation over channels. In addition, in a non-cooperative case, this Dec-POMDP will no longer be suitable.

Many existing works assume that the observation information obtained from an agent's neighbors is highly correlated. It can improve the efficiency of multi-agent coordination. In this case, exchange of local observations becomes important in coordination. From this view, we present a decentralized cooperative game model in which agents can iteratively adapt their strategies in terms of reduced competition or conflict, and can meet the minimum communication requirements for each agent timely. This presents a novel approach addressing the gaps in the aforementioned works.

## System Model and Problem Statement

In this section, we follow the basic idea of continuous time Markov model to define the basic model of a multi-Channel access problem, and then describe the specific functional definition of each variable and the decision model.

### Multi-Channel Access Model

We consider a multi-agent Ad-hoc network as being created by agents themselves in an open environment, with set $\mathcal{R} = \{r_1, r_2, ....r_N\}$ consisting of $N$ distributed agents. Although the multi-hops information sharing method can make each agent finally gain full knowledge of the global state, this consumes a lot of communication resources and also deteriorates the system's performance. Therefore, an agent makes decisions based on its limited observations, and the entire system would still be partially observed. The network consists of a set of contiguous, orthogonal (non-interfering) and homogeneous channels (e.g., 3 such channels in IEEE 802.11b/g and 12 in IEEE 802.11a), denoted by $CH = \{c_1, c_2, \ldots, c_K\}$. The available channels are also numbered from 1 to $K$, and we assume that $N > K$ agents are seeking channel opportunities in these $K$ channels.

We should recognize that agents can only access channels if the sensed channels are idle. As shown in Fig 1, a time slot consists of 3 parts: sensing, transmission and acknowledgment. Because of practical considerations, agent $r_i$ can sense a set of channels and a subset of sensed channels to access. Limited by its hardware constraints, $r_i$ can sense $\{C_1\}$ channels, ($\{C_1\} \in \{CH\}$, $|C_1| < K$) channels and access $\{C_2\}$ channels, ($\{C_2\} \in \{C_1\}$, $|C_2| < |C_1|$) channels. State statistics of the $K$ channels follows a discrete-time Markov process with $2^K$ states, where state is either *idle* or *occupied*. The channel sensing and access decisions are made to maximize agents reward by fully exploiting the sensing of vacant opportunities and the history statistics.

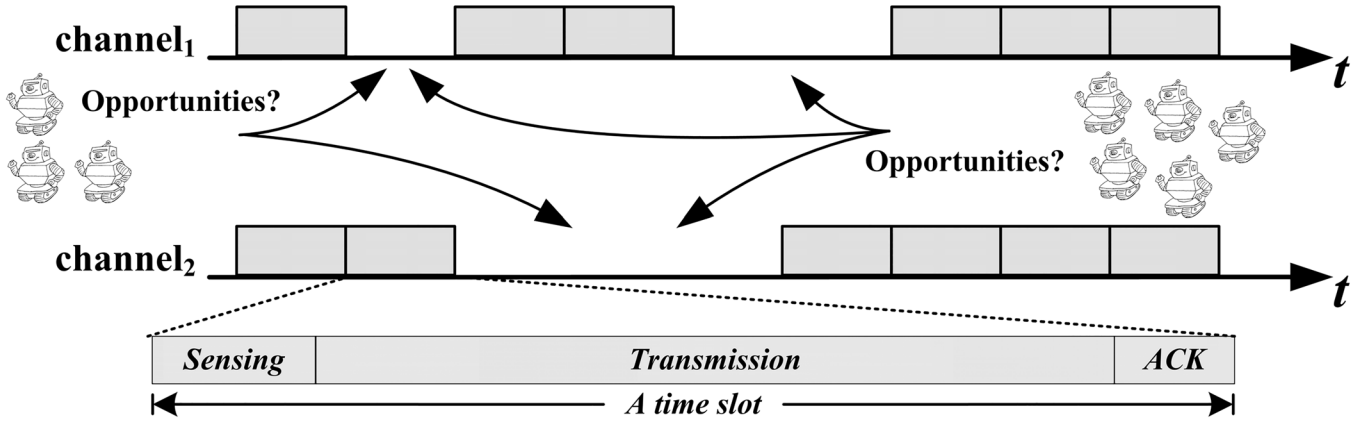


**Fig 1. Multi-agent Multi-channel Access Opportunities.** Several agents are independently seeking available communication opportunities in two channels, a suitable mechanism is required to ensure smooth communication and low conflicts.

doi:10.1371/journal.pone.0145526.g001

## Multi-channel Access Decision Problem

For some reasons, some agents change the status of stable channel access(switch to other channels, increasing the data flow or strong interference, etc.), and other agents need to adjust the channel access based on their limited observation so as to ensure the global rational use of resources and QoS. Therefore, the multi-agent multi-channel access problem can be described as cooperative searching for available resources in a partially observable multi-channel network. As such, this can be modeled as a Dec-POMDP problem in terms of interdependence. A finite-horizon Dec-POMDP can be defined as a tuple $< S, \Lambda, T, \mathbf{\Omega}, O, \mathbb{R}, p_0 >$, where

- $S = \{s_0, s_1, \ldots s_n\}$ denotes the finite set of network states.
- $A_i = \{a_1, a_2, \ldots, a_m\}$ denotes $r_i$'s available actions set. At each time step, all the agents in $\mathcal{R}$ take a joint action $\Lambda^t = \times_{1 \le i \le m}\{a_i^t\}$.
- $T$ denotes Markovian state transition function. $P(s'|s, \Lambda^t)$ denotes the probability that doing action $\Lambda^t$ and being in state $s$ then going to state $s'$.
- $\mathbf{\Omega}^t = \times_{1 \le i \le k}\{\omega_i^t\}$ denotes the set of joint observations of all the agents and $\omega_i^t$ is the observation by $r_i$ at time $t$.
- $O$ denotes observation function, which specifies the probability of joint observation $O(\mathbf{\Omega}^t|s', \Lambda^t)$.
- $\mathbb{R}(s'|\Lambda^t, s)$ denotes the reward value obtained from taking action $\Lambda^t$ in state $s$.
- $p_0 = \{B^0, s_0\}$ is the initial belief and state distribution.

Action $a$ is determined by the policy $\pi$: $b \rightarrow a$, which is the function that maps a belief state to the action that an agent should execute. $\mathbf{\Omega}_i^{t-1} = \times_{1 \le i \le t-1}\{\omega^{t-1}\}$ denotes the known network states.

Formally, most policies can be represented as decision trees. We use $Q_i$ to denote the possible policy space for agent $r_i$, and $Q_{-i}$ denotes the sets of policy trees for all agents except $r_i$. With a programming approach, it is required that we generate incrementally the sets of useful policies for each agent. Thus, a joint policy $\Pi = \times_{i \in N}\{\pi_i\}$ is a vector of policy trees. Evaluating a joint policy can then use the following formulation:

$$V(s, \Pi) = \sum_{\omega \in \Omega} P(\omega|s, \Pi)[\sum_{s' \in S} P(s'|s, \pi, \omega)V(s', \Pi(\omega))] \tag{1}$$

where $\Pi(\omega)$ is the joint policy of subtree selected after observation $\omega$. So we get the utility function as:

$$U(b_i, \Pi_i) = \sum_{s \in S} \sum_{\Pi_{-i} \in Q_{-i}} b_i(s, \Pi_{-i})V[s, \{\Pi_{-i}, \pi_i\}] \tag{2}$$

Therefore, the essence of this framework is to find a set of $n$ policies to maximize a total reward function from finite horizon $T$ under initial belief state $p_0$, and the expected joint reward is given by $E(\sum_{t=0}^{T} R(s^t, \Lambda^t)|p_0)$.

# A Resource-aware Approach for Multi-agent Multi-channel Access

In this section, we demonstrate an agent's decision-making process based on current observation and resource perception, and analyze the computational complexity under the instincts of *no-information-sharing*.

## Resource Awareness Policy Generation

From the idealistic view of the *Shannon's* theory [22], the optimal available resources under an ideal state for $r_i$ is:

$$Cap_i = B_n \sum_{j=1}^{N-1} \{(1-p_m) log_2(1+\frac{g_i^2 p_j^o}{\sigma_j^o}) + p_m log_2(1+\frac{g_i^2 p_j^o}{\sigma_i^p + g_i^2 p_i^p})\} \quad (3)$$

where $B_n$ is the channel bandwidth and $p_m$ is the channel state misperception probability. $\sigma_j^o$ and $\sigma_i^p$ respectively denote the noise variances from other agents and $r_i$ affected channel $ch_i$. $p_j^o$ and $p_i^p$ respectively denote the communication power of other agents and $r_i$. $g_i$ is the channel sensing gain. However, in the presence of sensing error, not only the sensing and access policy but also the operating characteristics of the channel sensor affect the performance of the network and the interference perceived by all the agents. The loss of resources caused by interferences are:

$$\triangle Cap_i = B_n \sum_{j=1}^{N-1} [log_2(1+\frac{g_i^2 p_j^o}{\sigma_j^o}) - log_2(1+\frac{g_i^2 p_j^o}{\sigma_i^p + g_i^2 p_i^p})] p_m \quad (4)$$

As a result, agent $r_i$ can obtain the idealistic expectation channel resources in $C_2$ as:

$$ECap_i = \sum_{i=1}^{|C_2|} (Cap_i - \triangle Cap_i) \quad (5)$$

We can see that the agent can access the network interval sequence independently, and this follows the same negative exponential distribution $G(t) = 1-e^{-\mu_i t}$, where $\mu_i$ is the channel free probability. Thus, we can get the probability of agent $r_i$ to choose and access channel $c_j$ as:

$$p_{i,j}^p = p_{i,j}^s V(ECap_i, \Pi_i) log_2(1+\frac{v_{i,j}}{ln\frac{0.2}{BER_{i,n}}}) \quad (6)$$

We can use $p_{i,j}^p$ and $p_{i,j}^s$ to denote the probability that agent $r_i$ select channel $c_j$ and the probability of channel $c_j$ being sensed idle respectively. $v_{i,j} = \frac{ECap_j \times p_j}{E(\triangle Cap_j)+N_0}$ is the Signal to Interference plus Noise Ratio (SINR) for agent $r_i$ from the other agents in channel $c_j$. This problem cannot be solved in one stage, and as such, should be done in an iterative manner. Therefore, based on the above analysis, we use Eqs (5) and (6) to obtain the policy tree and a target BER equal to $BER_{i,n} \approx \sigma_1 exp[\frac{-\sigma_2 \zeta_{i,n}}{2^{b_{i,n}}-1}]$, where $\sigma_1$ and $\sigma_2$ are Lagrangian multipliers, $b_{i,\,n}$ is the number of bits per symbol in channel $c_n$, and $\zeta_{i,\,n}$ is the Signal to Noise Ratio (SNR) for the receiver agent $r_i$ in channel $c_n$. Consequently, we adopt the utility function design in [23]:

$$U(b,\pi) = \mu_1 \sum_{c_i \in C_2} p_{i,j}^p log(\mu_2 c_i^h k_i) - cost_i \quad (7)$$

where the product $c_i^h k_i$ is the bandwidth (i.e. transmission rate), $c_i^h$ is the size of access channel in Hz, $k_i$ is the spectral efficiency[2] in bits per symbol per Hz due to adaptive modulation, and $\mu_1$ and $\mu_2$ are constants that depend on the communications protocol and agent communication system performance, respectively. $cost_i$ is the communication consumption, which relates to the agent's hardware system. The optimal policy is therefore $\pi^* = argmax[U(b, \pi_i)]$.

## Dynamic Local Search

Based on the model described above, agent $r_i$ cannot get a full view of the state of the system, since it can only use its observation to update its actions. The goal of this problem's model is to come up with a joint policy $\Pi^* = \times_{1 \le i \le N}\{\pi_i^t\}$, which can maximize the expected reward of all the agents over a finite horizon. The belief space is a sufficient statistic [21], and can be independent of the decision time. We remark that $r_i$ can only infer what action its neighbors may take, but the inference or conflict is inevitable. At each time slot, we can compute the expected value of a policy as follows:

$$E(V_\pi^t(\Omega^{t-1}, \omega^t)) = R(\Omega^{t-1}, \langle \omega^t, \pi^t \rangle) + \sum_{S' \in S} P(\Omega^{t-1}, \langle \omega^t, \pi^t \rangle, S') \cdot \sum_{\omega \in \Omega} O(S', \langle \omega^t, \pi^t \rangle, \omega') \cdot V_\pi^{t+1}(S', \omega') \tag{8}$$

Solutions to a finite-horizon POMDP can be represented as a decision tree, where nodes denote the actions and arcs denote the observations. Similarly, solving a finite-horizon Dec-POMDP with known state space can be formulated as a multiple vector of horizon $T$ policy tree searching process.

**Algorithm 1:** Resource aware policy search for agent $r_i$.

**Require:**
Set $g_0 = 0$; $\Re_0 = \{\Phi\}$ ; $\Re \in Q_i$;
**Ensure:**
$\exists\ \Pi^* = \times_{i \in N}\pi^*$ *and* $\forall\ v(\pi_i) \le v(\pi_i^*)$;
**for** *each* $r_i$ **do**
  *random select ploicy candidates set* $\{\eta_i\}$ *from* $\Re$;
  $g_i(t) = \max_{k \in (i \cup N_i)} g_k$;
  **for** *all* $\pi_i \in \eta_i$ **do**
    *excute* $\pi_i$ *to obtain* $\omega_t$;
    *compute* $g_i(t)$;
    **if** $g_i(t) > \hat{V}(\pi_0)$ **then**
      $\pi^* = \pi_i$;
    **else**
      *prune* $\pi_i$ *and get new* $\pi_i'$;
    **end if**
    $\eta_i' = argmax\{\Pi' \in \Re | V(\eta') > R(\eta)\}$ ;
  **end for**
  *return* $\pi^* \to \Pi^*$;
**end for**

As shown in Algorithm 1, $\Re \in Q_i$ denotes the random initialized policy space with completely unspecified candidate policies. $g_i = EV(\pi_i^*) - V(\pi_i^t)$ is the difference in value between the expected policy and the current one. In the beginning of each searching round, randomly select $\eta_0$ from $\Re$, if $g_i$'s value is bigger than $\hat{V}(\pi_0)$, then map $\pi_i$ to $\pi_i^*$. If not, prune the inappropriate $\pi_i$ and search new $\pi_i'$. We assume that the partial policy with the highest heuristic value is selected, and the provided value of $\hat{V}(\pi_0)$ is the *lower bound* for an optimal joint policy, which can be used to prune the search space. If $r_i$ has the minimum $g_i$ value in one round, then it will get priority to access its $\{C_2\}$. Other agents are constantly updated to the new strategy, and after finite times evaluation and exploration that they can get all the apposite policies to fix the $g_i$ value. In a limited belief space, by retrieving the limited policies space, and the state transition probability approaching the optimal values, similarly, the decision can approach the optimal policy $\Pi_i^*$. At each time slot, the computation of $g_i$ performs a summation over all possible network states and observations, and so the time complexity of this

algorithm is $O((|S|\cdot|Q_i|)^T)$. The value of a policy is highly dependent on the other agents' beliefs and the current system status, whereas, without sharing, the policy regeneration can only be derived on the basis of the reckoned joint policies. We define the policy update function as:

$$\begin{cases} \pi_i' = argmax\Xi_{\Pi}[R(\pi_i'|\delta_i^t, V(\pi^t))] \\ V(\pi^t) = \frac{1}{|C_2|}\sum_{1}^{|C_2|} R(ECap_i|a^t, \pi^t, \delta_i^t); \end{cases} \tag{9}$$

where $\Xi_\pi$ represents the conditional expectation given that policy $\pi_i$ is employed, and $B_0$ is the initial belief, which can be the stationary distribution of the network state. $\delta_i^t$ is the knowledge, consisting of two parts: channels observation $\omega^t$ and the known status $\Omega_i^t$. The search strategy performs a summation over all possible network states from agents' observations. Since each policy specifies different actions over possible histories of observations, the number of possible policies for agent $r_i$ is $O(|A_i|^{\frac{|\Omega|^T-1}{|Q_i|-1}})$. In consequence, the time complexity of finding the optimal policy by searching this space is: $O(|A_i|^{\frac{|\Omega|^T-1}{|Q_i|-1}}\cdot|S|\cdot|Q_i|)^T)$.

## A Decision Theoretical Approach for Multi-channel Access

In the previous sections, we proposed a random searching solution without coordination. This method has very high computational complexity and time cost. But from a practical point of view, each agent can be aware of its neighbor. Therefore, with the neighbor's policy sharing, the agent $r_i$ can get a proximate full local observation. Consequently, we refer to the design in [21], and the multi-agent finite horizon Dec-POMDP model can decompose into several single-agent POMDP decision problems.

### Neighbor-Aware Policy Generation

In order to solve a single-agent POMDP, we introduce neighbor policies $\bar{\pi}_i$ as a new parameter to the knowledge $\delta_i$, and the joint policy of $n$ neighbors is formulated as $\bar{\Pi}_i = \times_{i\in n}\{\bar{\pi}_i\}$. Therefore, we augment the state space to be $\Im = \{S \times \bar{S}\}$, where the second set $\bar{S}$ is the state variables of the other agents' beliefs. In consequence, we resolve and upgrade the Dec-POMDP to a POMDP model as a tuple $< \Im, A, T, \Omega, O, \mathbb{R}, \{\delta_i\} >$. All variable definitions remain unchanged, and to accomplish this, we factor the transition distribution into two terms: $T[(s', \bar{s}')|a, \bar{\Pi}_i(\bar{s}), (s, \bar{s})] = T[s'|a, \bar{\Pi}_i(\bar{s}), \bar{T}(\bar{s}'|s', a, \bar{\Pi}_i(\bar{s})]$, and the *upper bound* of the POMDP value function can be reached through the complete observation. In consequence, the belief update function can be denoted as:

$$b(s') = P(s'|\omega, a, b) = \frac{O(s', a, \omega)\sum_{s\in\Im} T(s, a, s')b(s)}{P(\omega|a, b)} \tag{10}$$

The value function of a POMDP is defined over the space of beliefs, where a belief state $b$ represents a probability distribution over states. The optimal value of policy $\pi^*$ can then be approximated as:

$$V_{\pi^*}(b) = \max_{a\in A}\{R_\delta(b, a) + \lambda\sum_{\omega\in\Omega} p(\omega|b, a)\cdot V^*(b, a, \omega)\} \tag{11}$$

## Heuristic Local Policy Search

Modeling our problem as a POMDP model is to search for the optimal policy $\pi^*$, and maximize the expected reward over a finite horizon-$T$ policy distribution over states. Formally, a belief state $b_{t+1} = P(s_{t+1}|\delta^t, \delta^{t-1}, \ldots, \delta^0)$ is a probability distribution over states conditioned on knowledge $\delta_i^t$. In order to avoid a heuristic with unbounded input (the knowledge can be arbitrary), a traditional approach is to learn a mapping from belief states to actions, which is from the known knowledge $\delta_i^t$. But in discrete worlds, beliefs can only be represented by a state with probabilities. We represent the regeneration process of belief states by sampling. A sample $x$ is annotated with a numerical importance factor to account for the difference in the sampling distribution.

Heuristic search is based on the decomposition of the evaluation function into a sequence of exact sub-evaluations. As aforementioned, we denote $q^t$ as an arbitrary depth $t$ policy vector extract from policy vector $Q^T$, and $\{q^t, Q^{T-t}\}$ constitutes a complete policy vector of depth $T$. This allows us to decompose the policy vector into any $t$ depth vector, and the value of the completion is:

$$V(Q^{T-t}|\{q^t, p_0\}) = V(p_0, q^t) + H^{T-t}(Q^{T-t}|q^t, p_0) \tag{12}$$

where $H(q)$ is the heuristic function, and the value of $Q^{T-t}$ depends on the previous execution and the underlying state distribution at time $t$. In consequence, we can describe the heuristic function as:

$$H^{T-t}(Q^{T-t}) = \sum_{s \in S} P(s|p_0, q^t) H^{T-t}(s) \tag{13}$$

As in Algorithm 2, randomly extract a sample $q^t$ from the possible policy space $Q$, and each node in the tree is a belief state $b_i$. For each encountered state $x_i$, belief state $b_i$ is updated to include the new state $x_i'$. In each sample searching, the agent selects the policy $b'$ at the greatest value. The sampling path terminates when it reaches a sufficient depth of the bounds of $T_q$, and goes back to the root so as to improve the *upper* and *lower bound* estimates. The search moves towards $\pi^*$ only with the acceptance probability $P(b^0)$, otherwise it remains at $b'$. At this point, the node $b^0$ becomes the root of the new search tree, and the remainder of the tree is pruned, as all other beliefs are now impossible. The search in new sample trees would not stop until there appears a policy to meet the resource requirements. Obviously, under a statistical hypothesis, the searching process converges to the expected distribution at a rate of $\frac{1}{\sqrt{H}}$, and $H$ denotes the sample size.

Algorithm 2: Sample extract-based search for agent $r_i$.

**Require:**
  random extract sample $\{q_i^t\}$ from $Q$; $v(b_0) = 0$;
**Ensure:**
  $\exists\ \forall\ v(\pi_i) \leq v(\pi_i^*)$;
1: random extract sample $\{q_i^t\}$ from $Q$;
2: **for** each $q_i^t$ **do**
3:  qualify $T_q$;
4:  **repeat**
5:  **for** each state $x_i$ from $b_i$ **do**
6:  compute $b(x_i')$, $x_i' \leftarrow T(x_i, a, x_i')$;
7:  **if** $b' \in q^t$ **then**
8:  continue to next $b_i$;
9:  **else**
10:  add $b'$ to $T_q$;
11:  **if** $U(b) < U(b')$ **then**

$b^0 = b'$;
**end if**
**end if**
*prune* $q^t$ *other than* $b^0$;
**end for**
*generate new* $q^{t\prime}$ *from root* $b^0$;
**until** $P(b^0) = min\left(1, \frac{b(x')T(x'|a,x)}{\sum_{x\in b} b(x)T(x'|a,x)}\right)$;
**end for**
*return* $\pi^*$;

A concise example is described in the following to illustrate our algorithmic process. We demonstrate a minimize-scale: two agents coordination. For each agent $r_i$, its action space has two actions {*Listen*, *Switch*}. These actions achieve channel perception, switch to other channels or stay in current channel, respectively. Each agent can sense two channels and choose one to access. The coordination has two states: establish a connection (R) or fail (W), denoted by $S = <R, W>$. The channel state misread probability is 0.3. The action-state transfer probability table as in Table 1, the initial joint action is $<S, S>$.

We define the highest reward (+50) to be the case when both agents get a good resource acquisition. A lower reward (-20) is agents' access in two different channels, and they can connect but with low resource acquisition. The worst case is lose connection (-100), and the cost of *Listen* is (-10). As shown in Fig 2, both agents start out with an initial belief state of b(s) = 0.5, and the discount factor is $\gamma = 0.9$. The first joint action at this belief is $<$ *Listen*, *Listen* $>$, the reward is (-20). As such, each agent has its own observation and network belief. In order to get a better reward, each agent removes all of the joint beliefs that are not consistent with its entire observation. After policies sharing, there is only a single possible belief $b(W) = 0.033$, and the optimal joint action for this belief is $<$ *Switch*, *Switch* $>$.

**Table 1. State-action transfer probability.**

| Action | S,S | L,S | S,L | L,L |
|---|---|---|---|---|
| **State** | | | | |
| W | 0.09 | 0.21 | 0.21 | 0.49 |
| R | 0.49 | 0.21 | 0.21 | 0.09 |

doi:10.1371/journal.pone.0145526.t001

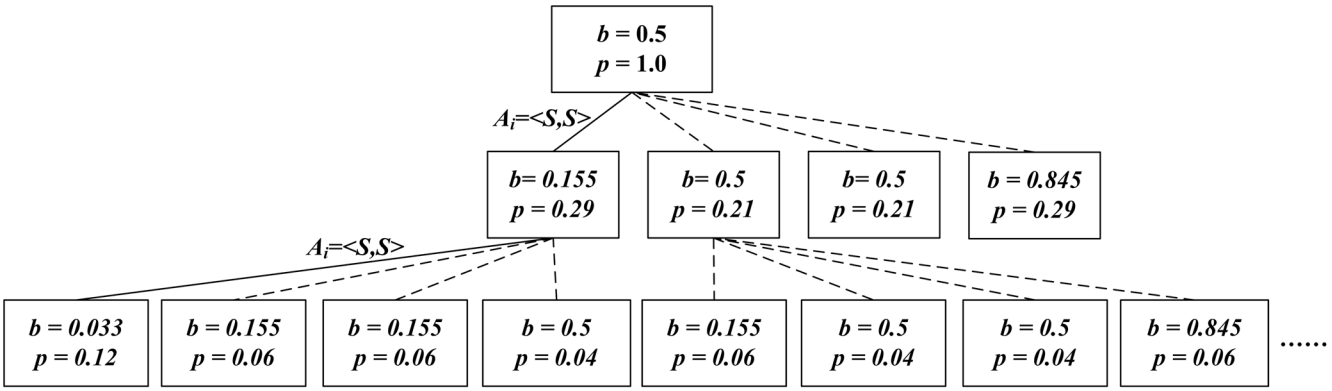


**Fig 2. Beliefs Update Processing.** A 3-step policy tree captured from Table 1, each of which can be conditioned on the outcome of previous actions. Each node is labeled with the action that should be taken if it is reached.

doi:10.1371/journal.pone.0145526.g002

It should be noted that there exists hidden competitions between agents for the finite resources (each agent wants to get more resources), that is, there should exist optimal joint policies to reach the Pareto optimal. But, it is infeasible for Dec-POMDP model because of the partial observation that we briefly described in the previous sections. In our design, it can finally reach the approximate Pareto optimal after a finite search. Therefore, it means that, in the finite belief space, there exists a pair of policies $\pi = (\pi_1, \pi_2)$ such that: $\forall_{\pi_1'}(V_1(\pi_1, \pi_2)) \geq V_1(\pi_1', \pi_2) \bigwedge \forall_{\pi_2'}(V_2(\pi_1, \pi_2)) \geq V_2(\pi_1, \pi_2')$. That is, for each agent, playing $\pi_i$ gives an equal or higher expected resource than playing $\pi_i'$. So both policies are best responses to each other.

## Implicit Competition Modeling and Equilibrium Analysis

As aforementioned, there exists hidden competition between agents for the finite channel resources, and techniques for eliminating dominated strategies in solving a POMDP are very closely related to techniques for eliminating dominated strategies in solving games in normal form [24]. From the game perspective, agents can get their locally optimal policy according to the Best Response (BR) dynamic iteration. In a general game, each agent negotiates and chooses the channels to maximize its payoff based on the channel situation in the last time slot observation, but the other agents (interference) can not change their channels simultaneously. However, BR does not guarantee convergence in all cases, and the stable state can not always be with the optimal overall reward. Hence, we study the characteristics of the multi-channel access game and its sub-optimal as in the following.

### Implicit Competition Game Model

According to the aforementioned, the access problem can be defined as a cooperative game $G = \langle \mathcal{R}, S, D_i, \mathbb{R} \rangle$, where the definitions of $\mathcal{R}$ and $S$ are unchanged, $D_i = \times_{1 \leq i \leq k}\{\pi_i\}$ is the finite set of policies available to agent $r_i$, $\mathbb{R}$ denotes the reward. We use $\theta_i(\pi_i)$ to denote the probability distribution assignment over policies available to agent $r_i$. Since agents select their policies simultaneously, agent $r_i$'s belief about the other agents' likely policies can be denoted as $\theta_{-i}$. If we define $V_{\pi_i}(s, \theta_{-i}) = \Sigma_{d_{-i}} \theta_{-i}(d_{-i}) V_i(\pi_i, d_{-i})$, then $B_i(\theta_{-i}) = \{\pi_i \in D_i | V_i(\pi_i, \theta_v) \geq V_i(\pi_i', \theta_{-i})\}$ denotes the best response function of agent $r_i$, which is the set of policies for agent $r_i$ that maximize its value of some belief about the policies of the other agents $d_{-i}$. Any policy that is not a best response to some belief can be abandoned.

Algorithm 3: General framework of competition equilibrium.

**Require:**
$\exists\, \theta_{-i} = \{b_1, \ldots, b_{i-1}, b_{i+1}, \ldots\}$ ;
**Ensure:**
$\forall\, s \in S$ and $v(\pi_i) \leq v(\pi_i^*)$;
**for** *each episode* **do**
  *Initialize get state S and D*;
  **repeat**
    *compute* $V_{\pi_i, d_{-i}}(b_i) \leftarrow O(s_i, a, s_i')$;
    **if** $V_{\pi_i, d_{-i}}(b_i) < \mathbb{R}_{min}^i$ **then**
      *prune* $\pi_i$ *and get new* $\pi_i'$;
    **else**
      *return* $\pi_i$ *to* $D'$;
    **end if**
  **until** $U_i(b_i) = \max \sum_{\pi_i \in D'} b_i(s, d_{-i}) V_{\pi_i}(s, \theta_{-i})$;
**end for**
*return* $\pi_i \rightarrow \pi^*$;

As in Algorithm 3, in a cooperative game, the reward functions for the game correspond to the reward functions of the POMDP, and an agent's belief is a distribution synthesization over the possible policies of the other agents. For each agent $r_i$, a belief is defined as a distribution over $S \times D_{-i}$, where the distribution is still denoted by $b_i$, and the utility of $b_i$ is:

$$U_i(b_i) = \max \sum_{\pi_i \in D'} b_i(s, d_{-i}) V_{\pi_i}(s, \theta_{-i}) \tag{14}$$

Given the set of policies and the reward function for a horizon-$t$'s game, the sets of policies and value functions for the $t$ horizon game are constructed by exhaustive backup. When a horizon-$t$'s POMDP is represented in the normal form with implicit competition, the policy sets include all depth-$t$ policy trees. Each policy profile is associated with a belief vector $\mathcal{B}$, reresenting the expected $t$-step cumulative reward achieved for each potential start state by following an apposite joint policy, whiles the size of the policy set for each agent $r_i$ is more than $A_i^{|O|^t}$, which is doubly exponential in the horizon-$t$. Because of the large sizes of the candidate policy sets, it is usually not feasible working directly. The search algorithm (Algorithm 2) we present in this paper only partially alleviates this problem by performing iterative elimination of dominated policies at each stage in the construction of the normal form representation, rather than waiting until the construction completes. Considering an $N$-player implicit competition game, we can formulate the game subject as:

$$\begin{cases} \sum_{n=1}^{K} \omega ln\left(1 + \dfrac{V(b_{i,n}) G_{i,i,n} \sigma_3}{\delta_i^2 + \sum_{j=1}^{K} p_{j,n} g_{j,i,n}}\right) - \mathbb{R}_{min}^i \geq 0 \\ \sum_{n=1}^{K} \omega ln\left(1 + \dfrac{V(b_{i,n}) G_{i,i,n} \sigma_3}{\delta_i^2 + \sum_{j=1, j \neq i}^{K} p_{j,n} g_{j,i,n}}\right) - \mathbb{R}_{exp}^i \geq 0 \end{cases} \tag{15}$$

In this constraints equations, $r_i$'s desired reward is no less than $\mathbb{R}_{min}^i$, and this guarantees a minimum level of resources achieved by each agent. $v(b_{i,\,n})$ is value of the belief distribution of agent $r_i$'s access in channel $c_n$, $\delta_i^2$ is the variance of the white Gaussian noise, and $\sigma_3 = \frac{\sigma_2}{ln\sigma_1}$. $\mathbb{R}_{exp}^i$ is the expected resource reward, and $G_{i,\,j,\,n}$ is the channel gain between two agents in channel $c_n$, and all policies should meet $\sum_{i=1}^{K} v(b_{i,n}) \leq \mathbb{R}_{\pi^*}(b, K)$. The existence and stability of the competition will be investigated in the following subsection.

## Evolutionary Equilibrium Analysis with Replicator Dynamics

In a multi-agent multi-channel access game, the stable state can be defined as the following: a joint policy $\Pi^*$ is and only if, for each agent and an arbitrary policy $\pi$ in its policy space, $v_i(\pi^*) \geq v_i(\pi, \theta_{-i})$ is always satisfied. Consequently, the process of this game can be modeled as a replicator dynamics, and this can be derived for each agent separately.

We consider a concise example with two new access agents $r_1$ and $r_2$. These agents appear first in the network with some spared channel opportunity (i.e., $c_1$ to $c_n$). With this specification, we analyze the evolutionary equilibrium for both deterministic and stochastic models. For the hidden competition among agents, the evolutionary equilibrium can be obtained as Replicator Dynamics solution [25], where $\chi_i$ denotes the proportion of the eager channel resources

that agents can get, and the replicator dynamics can be defined as the following:

$$\begin{aligned}\frac{\partial \chi_i^{b_i}(t)}{\partial t} &= \sigma \chi_i^{b_i}(t)[v_i^{b_i} - \bar{v}^{b_i}] \\ &= \sigma \chi_i^{b_i}(t)[U(\pi, \Pi_{-i}) - \bar{U}(\Pi_{-i})]\end{aligned} \tag{16}$$

where $\bar{v}^{(c_i)}$ is the estimated average reward for other agents in channel $c_i$, and the function $U$ is defined in Eq (7). With the two agents case, the evolutionary equilibrium is obtained as the solution of the following equation:

$$\begin{aligned}&\mu_1 log\ (\mu_2 \frac{\chi_1^{b_1} U(b_1)}{\chi_1^{b_1} U(b_1) + (1 - \chi_2^{b_2} U(b_2))}) - v_{\pi_1}(b_1) \\ &= \mu_1 log(\mu_2 \frac{\chi_2^{b_2} U(b_2)}{(1 - \chi_1^{b_1} U(b_1)) + \chi_2^{b_2} U(b_2)}) - v_{\pi_2}(b_2)\end{aligned} \tag{17}$$

where the terms on both sides of the equation are the rewards that the new access agents can get from their beliefs $b_1$ and $b_2$, respectively. Accordingly, the stability of the evolutionary equilibrium can be analyzed using the following Jacobian matrix:

$$\begin{bmatrix} \frac{\partial \sigma \chi_1^{b_1}[U(\pi, \Pi_{-1}) - \bar{U}(\Pi_{-1})]}{\partial \chi_i^{b_1}} & \frac{\partial \sigma \chi_1^{b_1}[U(\pi, \Pi_{-1}) - \bar{U}(\Pi_{-1})]}{\partial \chi_i^{b_2}} \\ \frac{\partial \sigma \chi_1^{b_2}[U(\pi, \Pi_{-2}) - \bar{U}(\Pi_{-2})]}{\partial \chi_i^{b_1}} & \frac{\partial \sigma \chi_1^{b_2}[U(\pi, \Pi_{-2}) - \bar{U}(\Pi_{-2})]}{\partial \chi_i^{b_2}} \end{bmatrix} = \begin{bmatrix} \mathcal{J}_{1,1} \mathcal{J}_{1,2} \\ \mathcal{J}_{2,1} \mathcal{J}_{2,2} \end{bmatrix} \tag{18}$$

where

$$\begin{aligned}\mathcal{J}_{1,1} = \quad & \sigma\{Z_2 - v_{\pi_1}(b_1) - \chi_1^{b_1}(Z_2 - v_{\pi_1}(b_1)) - (1 - \chi_1^{b_1}) \times [\mu_1 p_{1,2}^p log \frac{\mu_2 c_1^h k_1}{Z_1} - v_{\pi_2}(b_2)]\} \\ & -\sigma \chi_1^{b_1}\{\frac{\mu_1 U(b_1)}{\chi_1^{b_1} U(b_1) + (1 - \chi_2^{b_2} U(b_2))} + Z_2 - v_{\pi_1}(b_1) \\ & -\frac{\mu_1 \chi_1^{b_1} U(b_1)}{\chi_1^{b_1} U(b_1) + (1 - \chi_2^{b_2} U(b_2))} - \mu_1 p_{1,2}^p log \frac{\mu_2 c_2^h k_2}{Z_1} + v_{\pi_2}(b_2) + \frac{1 - \chi_1^{b_1} \mu_1 U(b_1)}{Z_1}\}\end{aligned} \tag{19}$$

$$\begin{aligned}\mathcal{J}_{1,2} = \quad & \sigma \chi_1^{b_1}\{-\frac{\mu_1 U(b_1)}{\chi_1^{b_1} U(b_1) + (1 - \chi_2^{b_2} U(b_2))} - \frac{(1 - \chi_1^{b_1}) \mu_1 U(b_1)}{Z_1} \\ & + \frac{\mu_1 \chi_1^{b_1} U(b_1)}{\chi_1^{b_1} U(b_1) + (1 - \chi_2^{b_2} U(b_2))}\}\end{aligned} \tag{20}$$

$$\mathcal{J}_{2,1} = \sigma\chi_1^{b_1}\{-\frac{\mu_1 U(b_1)}{\chi_1^{b_1}U(b_1)+(1-\chi_2^{b_2}U(b_2))} - \frac{(1-\chi_2^{b_2})\mu_1 U(b_1)}{Z_1} + \frac{\mu_1\chi_1^{b_1}U(b_1)}{\chi_1^{b_1}U(b_1)+(1-\chi_2^{b_2}U(b_2))}\} \tag{21}$$

$$\begin{aligned}\mathcal{J}_{2,2} = {} & \sigma\{Z_2 - v_{\pi_1}(b_1) - \chi_1^{b_1}(Z_2 - v_{\pi_1}(b_1)) - (1-\chi_1^{b_1}) \times [\mu_1 p_{1,2}^{p} log\frac{\mu_2 c_1^h k_1}{Z_1} - v_{\pi_2}(b_2)]\} \\ & -\sigma\chi_2^{b_2}\{\frac{\mu_1 U(b_2)}{\chi_1^{b_1}U(b_1)+(1-\chi_2^{b_2}U(b_2))} + Z_2 - v_{\pi_1}(b_1) \\ & -\frac{\mu_1\chi_2^{b_2}U(b_2)}{\chi_1^{b_1}U(b_1)+(1-\chi_2^{b_2}U(b_2))} - \mu_1 p_{1,2}^{p} log\frac{\mu_2 c_2^h k_2}{Z_1} + v_{\pi_2}(b_2) + \frac{1-\chi_2^{b_2}\mu_1 U(b_2)}{Z_1}\}\end{aligned} \tag{22}$$

where $Z_i$ specified as $Z_1 = (1-\chi_1^{b_1})U(b_1) + (1-\chi_2^{b_2})U(b_2)$ and $Z_2 = \mu_1 log\frac{\mu_2 c_1^h k_1}{\chi_1^{b_1}U(b_1)+\chi_2^{b_2}U(b_2)}$. The two eigenvalues of $\mathcal{J}$ can be obtained from $\Delta(\mathcal{J}) = \frac{\mathcal{J}_{1,1}+\mathcal{J}_{2,2}\pm\sqrt{4\mathcal{J}_{1,2}\mathcal{J}_{2,1}+(\mathcal{J}_{1,1}-\mathcal{J}_{2,2})^2}}{2}$, and the evolutionary equilibrium is stable if these two eigenvalues have negative real parts [23].

## Approximate Fair Maximization Policy Analysis

Among the different Cooperative Game solutions, it is important to note that the issue about fairness in this context, e.g., new access agents, is different from the case of resource occupation among the early-existing agents in the network. In this section we will analyze approximate fairness of the game, and discuss the feasibility of the proposed neighbor-aware channel access scenario in the previous section. In this scenario, the approximate Pareto optimal result can satisfy all agents' minimum requirements. Typically, if a channel is occupied, the other agents should be denied access to the frequency band.

In the proposed competitive game model, a virtual-feasible resource access assignment set is existent, hence, we can use a bounded set $\mathbb{F} = \{\mho_{min}^1, \mho_{min}^2, ..., \mho_{min}^{C_2}\}^T$, which denotes the minimum resource required of the game. The vector $r = \{\mathbb{R}_{exp}^1, \mathbb{R}_{exp}^2, ..., \mathbb{R}_{exp}^{C_2}\}^T$ represents the set of rewards for the access agents. The reward vector $r \in \mathbb{R}^{K+1}$ in the $K$ channels can form the fairness problem $\phi(\mathbb{F}, r)$. It has been shown that there exists a unique equilibrium, which can be calculated by Eq (17):

$$\phi(\mathbb{F}, r) = argmax\prod_{i=1}^{K}\mathbb{R}_{exp}^i - v\chi_i(\mho_i) \tag{23}$$

Hence, we can use Eq (23), to confirm the selected solution for stable point in the previous section. This is also the point where "egalitarian" solutions of the game come in, and one such method is applicable to the equal gains principle, a Pareto optimal. For the 2 agents case in the previous section, the proportion $\chi_i$ in $\phi$, which is weakly efficient and satisfies the equal gain condition $\chi_1(\mho_1) - r_1 = \chi_2(\mho_2) - r_2$, is called the "egalitarian" solution. As mentioned earlier, in our resource access method, the stable acts as a marketplace where the primary and secondary systems can do bargaining. The fair solution for two agents about one channel access is at

the intersection of the egalitarian solutions:

$$\begin{cases} U_1 - \chi_1(\mho_1) = U_2 - \mathbb{R}^2_{exp} \\ U_1 + U_2 = argmax_{\chi_i}(U'_1 + U'_2) \end{cases} \quad (24)$$

Condition Eq (24) dictates that the operating point should be on the boundary of the minimum region. Therefore, the intersection gives the unique approximate fairness solution. For a $N$ agents game, the fairness problem Eq (23), should be solved by calculating the $\chi_i$, $i = 1, 2, \ldots, N$. To verify the case, we note that the stable point in the proposed design is also on the perpendicular boundary to (24) at its intersection. The corresponding optimization satisfies Eqs (15) and (17), defined by $\max_{\chi_i}[U_1 - \chi_1(\mho_1)] \times [\mathbb{R}^2_{exp} - U_2]$, which is subject to $(r_i, \mathcal{B})$, $i = 1, 2, \ldots N$. It is straightforward to confirm the solution of Eq (18), as it satisfies the description at the beginning of this section.

## Experiments and Results

In this section, we designed several experiments to evaluate the proposed methods in above sections. We employed the multi-agent platform in [4] to simulate the multi-agent Ad-hoc network. The data unit length was fixed at 1,024 bytes. We evaluate the performance of the proposed scheme with wide band available by simulation and compare it with the priced-based centralized channel allocation method (OPTIMAL) [26] and the RANDOM method to validate the efficiency, which allows agent randomly accesses channels from its current belief on each channel. The simulation parameters are shown in Table 2.

In order to facilitate the numerical statistics, we use one channel for global listening (especially for the OPTIMAL method of the centralized resource allocation), the rest of 10 channels allow agents to access. We conducted the simulations under various scale agents and the simulation results are the average value of 100 runs.

### Resource Lost Rate

As in Fig 3, the results show that the influences of different channel access strategies, have direct impact on the available channel resources. The axes represent the number of agents and the resource loss rates. Agents adopt a RANDOM method, and with the expansion of scale (20–200, from 0 agent there has no conflict), the congestion and resource loss rates continue to rise closing to 90%. Meanwhile, in our algorithm, agents communicate with their neighbors to exchange decision policies, and acquire a better joint behavior through continuous negotiation and iterations (the average max amount of loss is 65.38%, the average sample variance is 8.12%), the variance shows that our algorithm is more stable.

**Table 2. Simulation parameters.**

| Simulation Parameters | Value |
|---|---|
| Number of agents | 200–1000 |
| Maximum number of channels | 11 |
| Number of perception channels | 5 |
| Number of access channels | 3 |
| Maximum resources for a channel | 100 |
| Data frame size | 1 |

doi:10.1371/journal.pone.0145526.t002

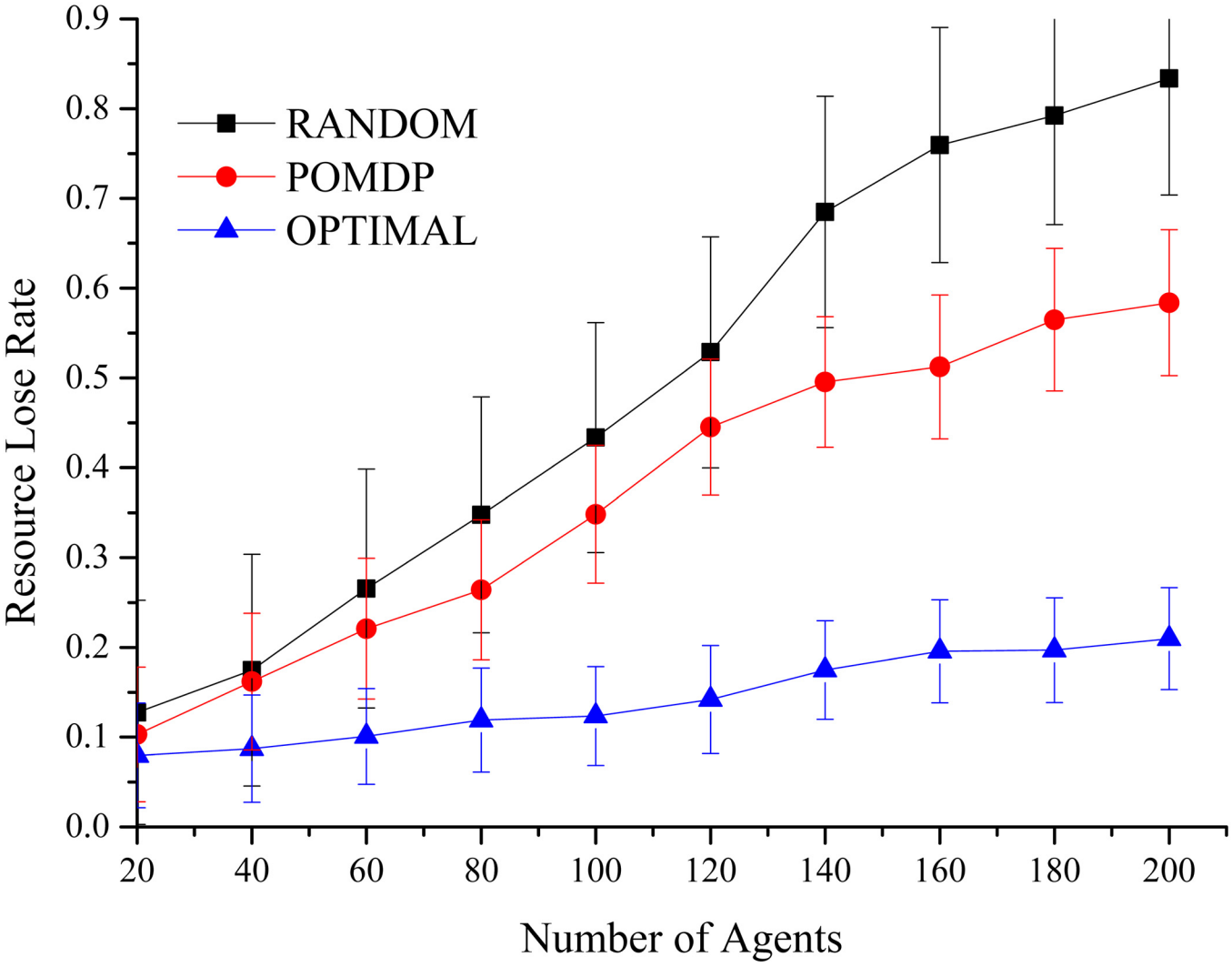


**Fig 3. Resources Loss Rate.** With the increased size of the agents, the randomness of the RANDOM method increased interference between agents, which brings down the network resources utilization. OPTIMAL method can maintain an efficient use of the resources, but its time consumption is much larger than the self-decision methods. POMDP methods maintain a relative balance to the above methods.

doi:10.1371/journal.pone.0145526.g003

The OPTIMAL method can provide a better result, but the resource consumption of global consultations could not be avoided (the average max amount of loss is 21.92%, the average sample variance is 5.67%). Furthermore, due to agents' misperception and accessing, the resource loss (conflicts) is inevitable, and will increase sharply with the expansion of the agents.

## Resource Available Rate

With the premise of partial observation, we set the RANDOM and our algorithm to start from the initial belief probability 0.5, as shown in Fig 4. But the difference is, our algorithm can reach an average resource showed at 52.67% and the average sample variance is 3.32%. RANDOM's resource obtain rapid descent, and when there has 200 agents, the available resources only remain 33.48%, but with 12.16% average sample variance. When the number of agents and network resources are relatively homogeneous, the available resources rate can approach

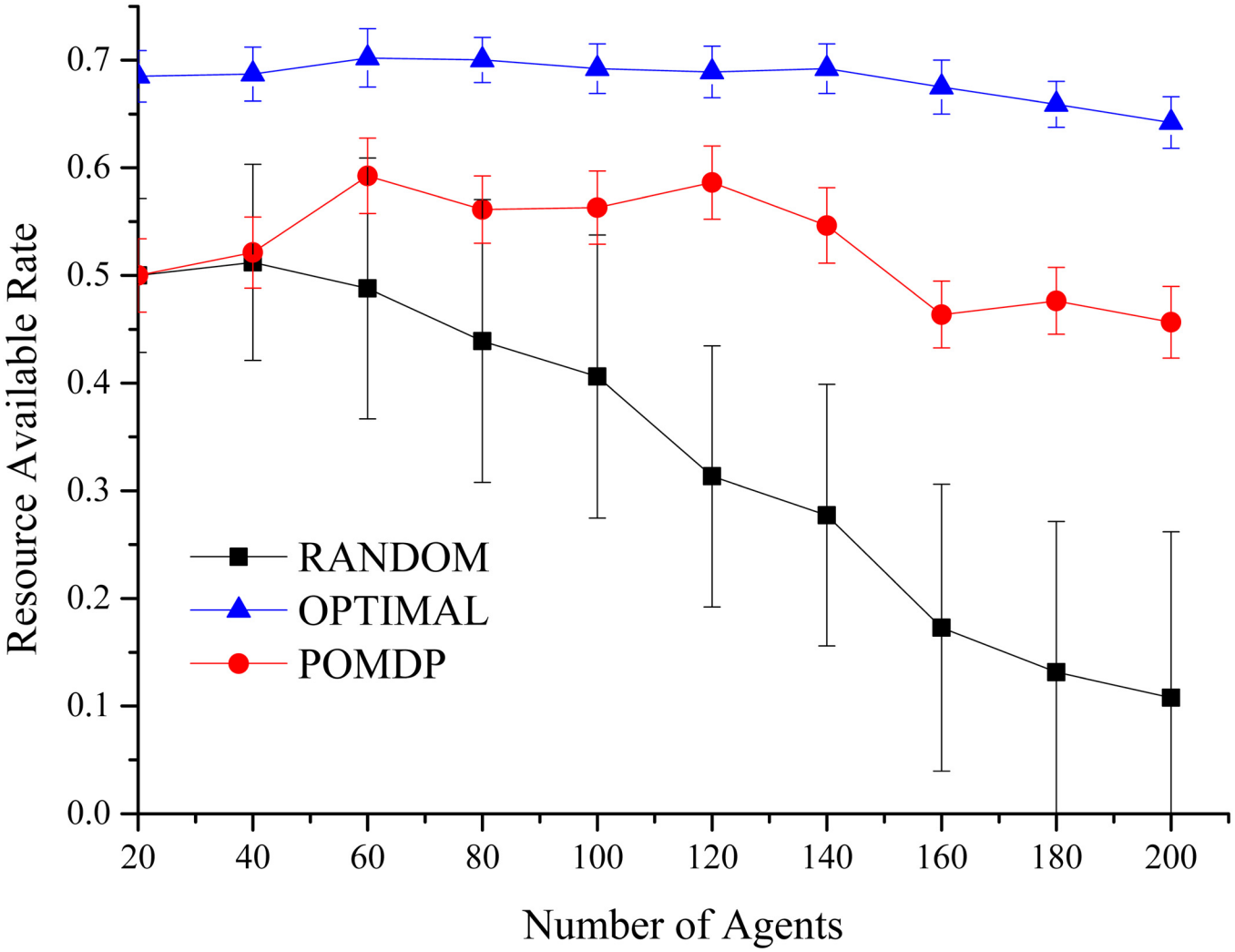


**Fig 4. Resources Available Rate.** Under the same experimental setup, with the increasing size of the agents, RANDOM method reduces the resources available for each agents. Because of neighbor's awareness in POMDP, the agents can be maintained in a state of relative balance (less variance than RANDOM).

doi:10.1371/journal.pone.0145526.g004

the expected value. Then due to the increase of the agents' number, the available rates decrease. The OPTIMAL can keep an average resource obtained at 68.23% and with a very stable average sample variance 2.37%.

Obviously, agents can obtain more resources with our design than RANDOM provides. Especially, with the increase of the agents' number and passing time (agents can exchange information with neighbors and accumulate from known knowledge), the resource availability rate remains in a relatively stable state until agents reaching network's saturation.

## Available Resource in Different Interaction Frequencies

In this simulation, we test the average available channel resources for the new accessed agents under different interaction frequencies of the other agents in the network. We set 5 channels and 100 per slot new accessed agents, which are uniformly distributed in these 5 channels, the max agents number is 1000. The interaction frequency of the other agents was set to $r$ = [0.2,

0.4, 0.6, 0.8]. X-axis represents the agents' number. As in Fig 5, we can find two significant changes for the new accessed agents: with increasing numbers of network agents or the higher interaction frequency, the available resources decline. In addition, when the number of agents is more than the maximum number the network can support, the available resources for the entire network will be sharply reduced.

According to the experiment's results, we can make a bold hypothesis that while the number of agents and the resource relatively balance, there should be a suitable interaction frequency that makes each agent obtain available resources to maximize its utility.

## Resource Available in Different Assignments

In this simulation, we discuss the relationship between different team sizes when agents access channels under different assignment. We divide 100 agents into different team sizes, and allow them to access 5 channels. Simulation results are shown in Fig 6. Caused by new access agent,

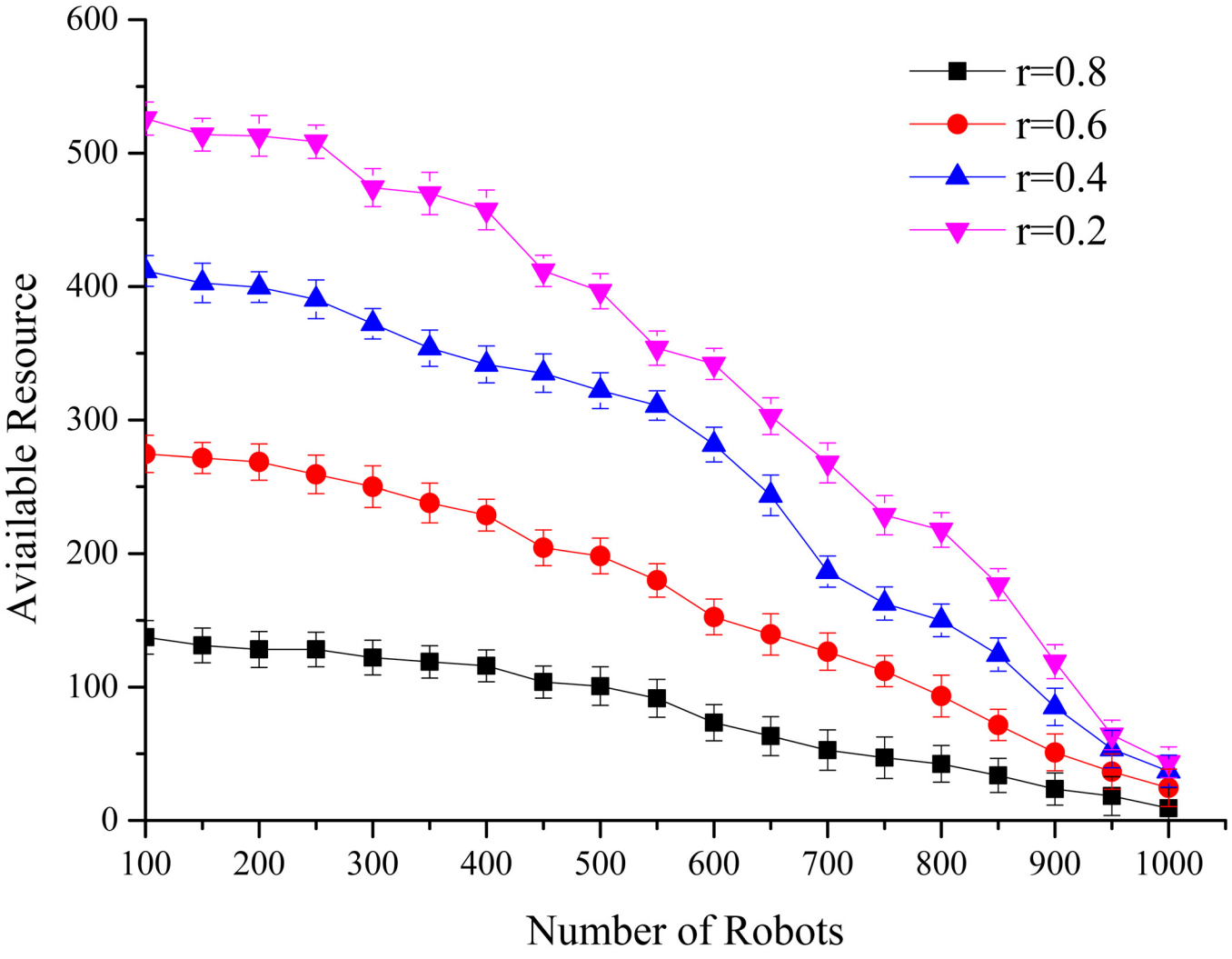


**Fig 5. Available Resources in Different Interaction Frequencies.** In different interaction frequencies, the available resources shrink with the increasing number of agents. Similar to the allocation of limited resources in human society, the average gain decreases with the increasing number of people. Experimental results are consistent with the general understanding.

doi:10.1371/journal.pone.0145526.g005

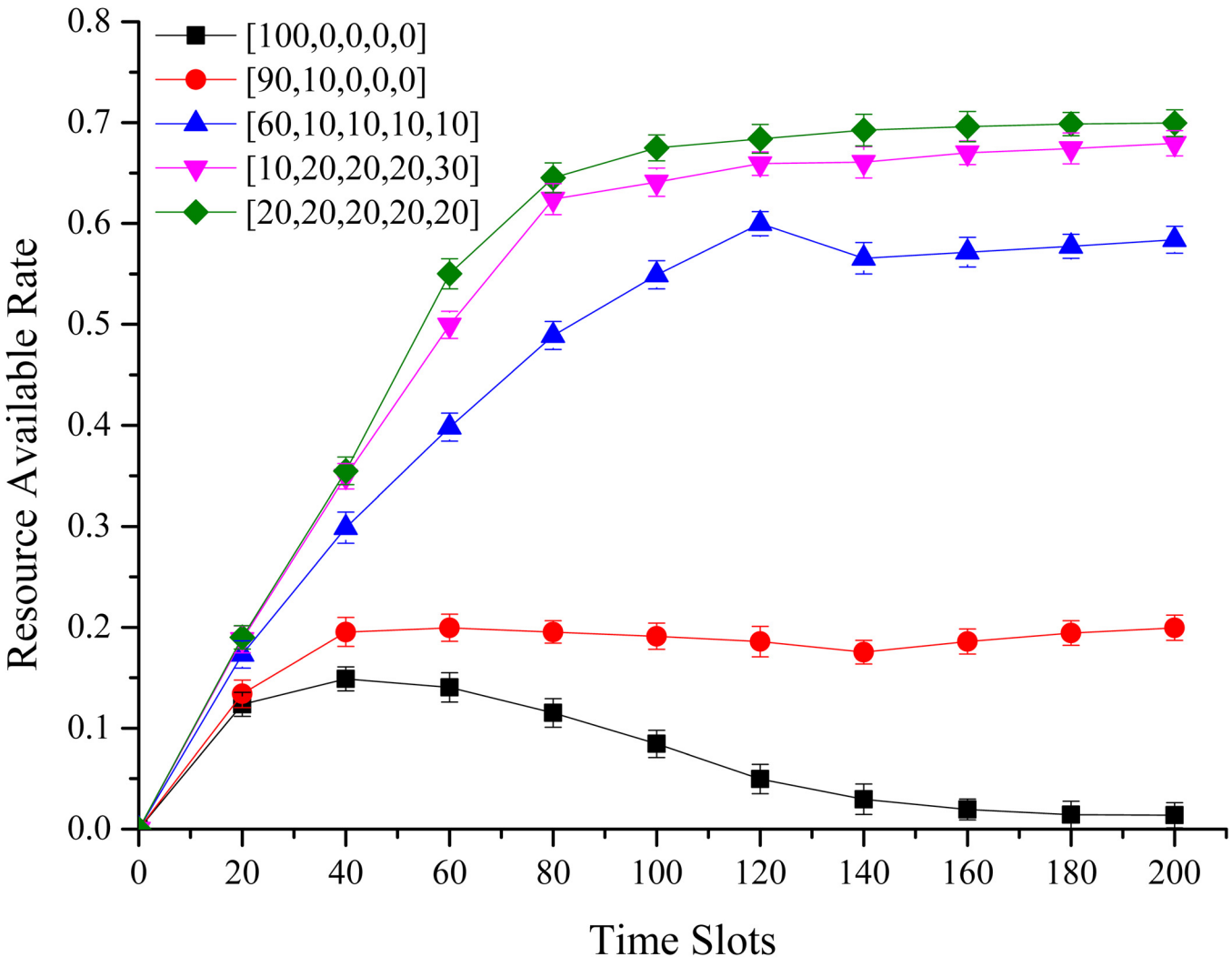


**Fig 6. Resource Available in Different Assignments.** In the 5 specified channel, the more uniform distribution of the agents, the higher the probability of their available resources, whereas the performance reduces (more crowded, no elimination of competition that makes the average income is lower).

doi:10.1371/journal.pone.0145526.g006

blocking rate of channel will increase and influence the original agents due to the partial observation.

Obviously, we can see that when the combination in each channel distribution is more uniform, the greater resource available, as assignment [10, 20, 20, 20, 30] and [20, 20, 20, 20, 20]. In an extreme access situation with [100, 0, 0, 0, 0], all agents are in one channel. When communication demand escalates, all agents almost have no chance to obtain available resources. From above analysis, we can conclude that agents will gain more available resources when they distribute more uniformly.

## Resource Available Comparison

Fig 7 displays the contrast of channel resources awareness between our algorithm and RANDOM. In this simulation, we set 100 agents with freedom interactive frequency.

It can be seen that, with 500 tests for the same channel (distribution within the circles), agents can obtain the actual state of the channel. The red trail denotes the search result by

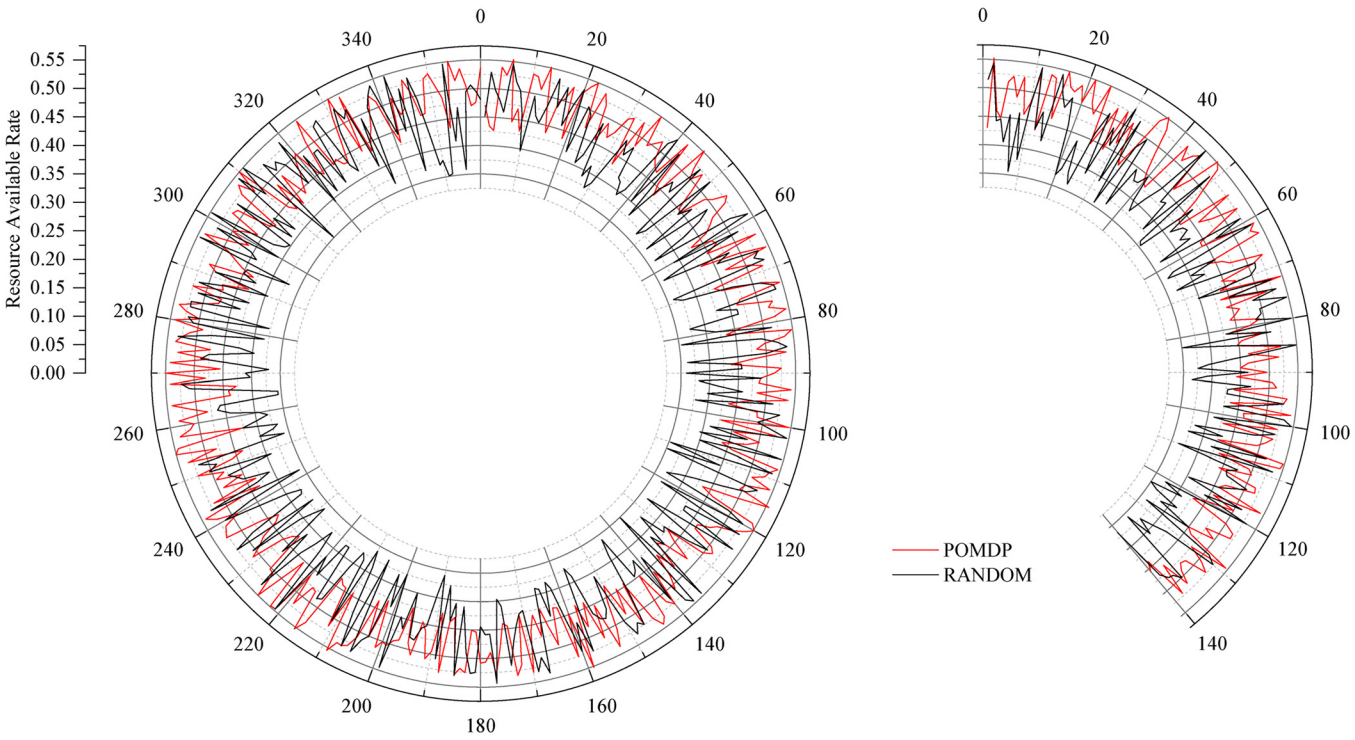


**Fig 7. Resource Perception of 500 Tests.** To illustrate the variances in the 4 aforementioned simulation results, this figure gives the resource perception comparison in 500 tests between POMDP and RANDOM.

doi:10.1371/journal.pone.0145526.g007

POMDP, and black trail denotes the RANDOM method. It is obvious that the randomness and divergence of RANDOM far outweigh that of POMDP.

## Conclusion

We assumed in this paper that channel state transition probabilities can be entirely perceived, but in practice, this may not be available. The problem then becomes a decision model with unknown transition probabilities, but such mode is beyond the scope of this paper. In our design, we reduce a Dec-POMDP model to a simplified one by separating the problem into single-agent decision coordination, which may result in a low-complexity but potentially suboptimal design. In practical applications, systems Dynamics making use of pure policy space searching to solve all the problems become impractical, and need to be adjusted according to the actual situation and dynamics, and add more factors. In our future work, we will pursue the optimal joint design of the tradeoff between complexity and optimality, and will apply reinforcement learning theory on real multi-robots platform.

## Supporting Information

**S1 Table. Experiment Data for Resource lost Rate Comparison.**
(XLS)

**S2 Table. Experiment Data for Resource Available Rate Comparison.**
(XLS)

**S3 Table. Experiment Data for Available Resource in Different Interaction Frequencies Comparison.**
(XLS)

**S4 Table. Experiment Data for Resource Available in Different Assignments Comparison.**
(XLS)

**S5 Table. Experiment Data for Available Resource Perception.**
(XLS)

**S1 File. Supplementary Methods and Datasets Introduction. Supporting Information Supplementary Methods and Datasets Introduction.**
(DOC)

## Acknowledgments

This research was sponsored by the NSFC 61370151 and 61202211, the National Science and Technology Major Project of China 2015ZX03003012, the Central University Basic Research Funds Foundation of China ZYGX2014J055, and the Science and Technology on Electronic Information Control Laboratory Project.

## Author Contributions

Conceived and designed the experiments: YX ML AWM. Performed the experiments: ML YX AWM. Analyzed the data: ML YX AWM. Contributed reagents/materials/analysis tools: YX ML AWM. Wrote the paper: ML AWM YX. Responsible for the theoretical study and mathematical derivation: ML. PI (Principal Investigator) of all the funding projects, and, as the first and third authors' PhD adviser, responded to problem modeling and algorithm layout: YX. Presided over the experimental test bed design and language revision: AWM.

## References

1. Wang T, Dang Q, Pan P. A Multi-Robot System Based on A Hybrid Communication Approach. Studies in Media and Communication. 2013; 1(1):91–100. doi: 10.11114/smc.v1i1.124
2. Iqbal, J, Yousaf, MM, Awais, MM, editors. A scalable approach of message interpretation by demonstrations for multi-robot communication. Multitopic Conference, 2009 INMIC 2009 IEEE 13th International; 2009: IEEE. 10.1109/INMIC.2009.5383082
3. Conti M, Giordano S. Mobile ad hoc networking: milestones, challenges, and new research directions. Communications Magazine, IEEE. 2014; 52(1):85–96. doi: 10.1109/MCOM.2014.6710069
4. Zhang Y, Xu Y, Hu H. Cooperative Decision Algorithm for Time Critical Assignment without Explicit Communication. Intelligent Information Processing VII: Springer; 2014. p. 197–206.
5. Liu M, Xu Y, Wu S, Lan T. Design and Optimization of Hierarchical Routing Protocol for 6LoWPAN. International Journal of Distributed Sensor Networks. 2015; 2015. doi: 10.1155/2015/802387
6. Ab Wahab MN, Nefti-Meziani S, Atyabi A. A Comprehensive Review of Swarm Optimization Algorithms. PLoS ONE. 2015; 10(5): e0122827. doi: 10.1371/journal.pone.0122827 PMID: 25992655
7. Zhao Q, Sadler BM. A survey of dynamic spectrum access. Signal Processing Magazine, IEEE. 2007; 24(3):79–89. doi: 10.1109/MSP.2007.361604
8. Kulkarni RV, Venayagamoorthy GK. Particle swarm optimization in wireless-sensor networks: A brief survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 2011; 41(2):262–7. doi: 10.1109/TSMCC.2010.2054080
9. Yan Z, Jouandeau N, Cherif AA. A survey and analysis of multi-robot coordination. International Journal of Advanced Robotic Systems. 2013; 10:399. doi: 10.5772/57313
10. Capitan J, Spaan MT, Merino L, Ollero A. Decentralized multi-robot cooperation with auctioned POMDPs. The International Journal of Robotics Research. 2013; 32(6):650–71. doi: 10.1177/0278364913483345

11. Tan L, Feng Z, Li W, Jing Z, Gulliver TA. Graph coloring based spectrum allocation for femtocell downlink interference mitigation. Wireless Communications and Networking Conference (WCNC), 2011 IEEE; 2011: IEEE. 10.1109/WCNC.2011.5779338.
12. Xu Y, Wang J, Wu Q, Anpalagan A, Yao Y-D. Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games. Selected Topics in Signal Processing, IEEE Journal of. 2012; 6(2):180–94. doi: 10.1109/JSTSP.2011.2176916
13. Tang S, Mark BL, editors. Performance analysis of a wireless network with opportunistic spectrum sharing. Global Telecommunications Conference, 2007 GLOBECOM'07 IEEE; 2007: IEEE. 10.1109/glocom.2007.880.
14. Liu H, Krishnamachari B, Zhao Q, editors. Cooperation and learning in multiuser opportunistic spectrum access. Communications Workshops, 2008 ICC Workshops' 08 IEEE International Conference on; 2008: IEEE. 10.1109/ICCW.2008.98.
15. Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 2008; 38 (2):156–72. doi: 10.1109/TSMCC.2007.913919
16. Fernandez-Gauna B, Graña M, Lopez-Guede JM, Etxeberria-Agiriano I, Ansoategui I. Reinforcement Learning endowed with safe veto policies to learn the control of Linked-Multicomponent Robotic Systems. Information Sciences. 2015; 317:25–47.
17. Fernandez-Gauna B, Etxeberria-Agiriano I, Graña M. Learning Multirobot Hose Transportation and Deployment by Distributed Round-Robin Q-Learning. PLoS ONE. 2015; 10(7): e0127129. doi: 10.1371/journal.pone.0127129 PMID: 26158587
18. Halldórsson MM, Halpern JY, Li LE, Mirrokni VS, editors. On spectrum sharing games. Proceedings of the twenty-third annual ACM symposium on Principles of distributed computing; 2004: ACM. 10.1145/1011767.1011783.
19. Yichen W, Pinyi R, Zhou S. A POMDP based distributed adaptive opportunistic spectrum access strategy for cognitive ad hoc networks. IEICE transactions on communications. 2011; 94(6):1621–4.
20. Seuken S, Zilberstein S. Improved memory-bounded dynamic programming for decentralized POMDPs. arXiv preprint arXiv:12065295. 2012.
21. Feng M, Qu H, Yi Z. Highest Degree Likelihood Search Algorithm Using a State Transition Matrix for Complex Networks. Circuits and Systems I: Regular Papers, IEEE Transactions on. 2014; 61 (10):2941–50. doi: 10.1109/TCSI.2014.2333677
22. Kish LB, Harmer GP, Abbott D. Information transfer rate of neurons: stochastic resonance of Shannon's information channel capacity. Fluctuation and Noise Letters. 2001; 1(01):L13–L9. doi: 10.1142/S0219477501000093
23. Niyato D, Hossain E, Han Z. Dynamics of multiple-seller and multiple-buyer spectrum trading in cognitive radio networks: A game-theoretic modeling approach. Mobile Computing, IEEE Transactions on. 2009; 8(8):1009–22. doi: 10.1109/TMC.2008.157
24. Hansen E A, Bernstein D S, Zilberstein S. Dynamic programming for partially observable stochastic games. AAAI. 2004; 4:709–715.
25. Roca CP, Cuesta JA, Sánchez A. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. Physics of life reviews. 2009; 6(4):208–49.
26. Xue Y, Li B, Nahrstedt K. Optimal resource allocation in wireless ad hoc networks: A price-based approach. Mobile Computing, IEEE Transactions on. 2006; 5(4):347–64. doi: 10.1109/TMC.2006.1599404