INVITED PAPER *Special Section on Exploring Drone for Mobile Sensing, Coverage and Communications: Theory and Applications*

Multi-Autonomous Robot Enhanced Ad-Hoc Network under Uncertain and Vulnerable Environment*

Ming FENG[†a)], **Lijun QIAN**[††b)], ***and*** **Hao XU**[†c)], ***Nonmembers***

SUMMARY This paper studies the problem of real-time routing in a multi-autonomous robot enhanced network at uncertain and vulnerable tactical edge. Recent network protocols, such as opportunistic mobile network routing protocols, engaged social network in communication network that can increase the interoperability by using social mobility and opportunistic carry and forward routing algorithms. However, in practical harsh environment such as a battlefield, the uncertainty of social mobility and complexity of vulnerable environment due to unpredictable physical and cyber-attacks from enemy, would seriously affect the effectiveness and practicality of these emerging network protocols. This paper presents a GT-SaRE-MANET (Game Theoretic Situation-aware Robot Enhanced Mobile Ad-hoc Network) routing protocol that adopt the online reinforcement learning technique to supervise the mobility of multi-robots as well as handle the uncertainty and potential physical and cyber attack at tactical edge. Firstly, a set of game theoretic mission oriented metrics has been introduced to describe the interrelation among network quality, multi-robot mobility as well as potential attacking activities. Then, a distributed multi-agent game theoretic reinforcement learning algorithm has been developed. It will not only optimize GT-SaRE-MANET routing protocol and the mobility of multi-robots online, but also effectively avoid the physical and/or cyber-attacks from enemy by using the game theoretic mission oriented metrics. The effectiveness of proposed design has been demonstrated through computer aided simulations and hardware experiments.

key words: *reinforcement learning, game theory, mobile ad-hoc network, mission oriented metrics, multi-agent systems*

1. Introduction

Mobile Ad-hoc network (MANET) is a self-configuring network of mobile routers connected by wireless links, i.e. the union of which form an arbitrary topology. The routers are able to move randomly and/or organize themselves arbitrary [1]. Therefore, the wireless network topology can be reconfigured rapidly [2], [3]. For instance, a network can be formed dynamically by the wireless nodes to exchange information without using any fixed existing network infrastructure [4]. Each node plays a role of router in the MANET as it must forward the traffic to other nodes. During the past decade, advanced MANET networking development in harsh environments, e.g. battlefield, outer space, disaster rescue and etc., attracts tremendous interests [5], [6]. The major challenge is how to provide reliable information exchange even while lacking continuous network connectivity due to uncertain and vulnerable environment. Delay Tolerant Networking (DTN) [5] is an emerging Ad-hoc network that can support multi-user with sporadic connectivity. Meanwhile, inspired by carry-and-forward [6] mechanism adopted in E-mail exchanging system, researchers have proposed a class of store-carry forward opportunistic mobile networking. Moreover, opportunistic mobile networking has been considered as another promising networking for uncertain environment. To reduce the network delay and improve the packet delivery ratio, how to plan the moving paths for network nodes especially under uncertain and vulnerable environment, e.g. at tactical edge, and when to forward messages are two critical factors for opportunistic mobile networking.

To address two important challenges in opportunistic mobile networking, a series of routing protocols have been developed recently. Epidemic routing proposed in [7] indiscriminately floods the network with messages. As shown in [7], this routing could provide a high message delivery ratio and delivery time. However, it will increase message delivery cost as well. For reducing the message delivery cost, many recent researches [8]–[10] were inspired by efficiently connectivity existing in social network. Through developing effective social metrics, social network has been successfully integrated into communication network as a new type of network. Particularly, three promising social-based routing protocols, i.e. SimBet [8], Bubble RAP [9] and Friendship routing [10], were developed and applied to uncertainty communication environment. In [8], SimBet used similarity and betweenness centrality metrics to determine the suitable relay nodes with higher probabilities of delivering the message. Bubble RAP [9] utilized centrality and community to make forwarding decisions whereas friendship routing [10] considered the interrelations among nodes through introducing a metric to measure the quality of friendship. However, those social-aware mobile opportunistic network protocols mainly focused on effectively introducing social functions (e.g. carry-and-forward scheme [6]) into network

Manuscript received December 25, 2018.
Manuscript revised March 9, 2019.
Manuscript publicized April 26, 2019.

[†]The authors are with University of Nevada, Reno, NV, USA.
[††]The author is with Prairie View A&M University, TX, USA.

[*]This research work is supported in part by the U.S. Office of the Under Secretary of Defense for Research and Engineering (OUSD (R&E)) under agreement number FA8750-15-2-0119. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of the Under Secretary of Defense for Research and Engineering (OUSD (R&E)) or the U.S. Government.

a) E-mail: mingf@nevada.unr.edu
b) E-mail: liqian@pvamu.edu
c) E-mail: haoxu@unr.edu

DOI: 10.1587/transcom.2018DRI0001

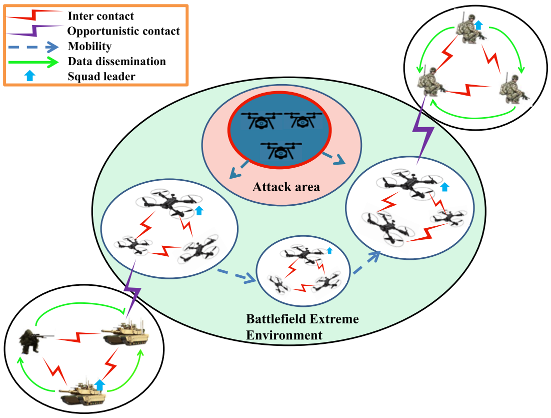


Fig. 1 Multi-robot enhanced MANET at tactical edge.

nodes without considering how to use social mobility further enhancing network interoperability,the ability for two or more networks, systems, devices, applications or components to communicate. Recently, authors in [11] developed a novel SCATE (Social-Cognitive Advancement at Tactical Edge) routing that effectively upgrade the MANET quality through utilizing learned social mobility. However, the effects from uncertain and vulnerable environment such as enemy attacks at tactical edge are not considered which could limit their applicability to real-world scenario such as battlefield.

Based on the discussion above and to further reap the advantages from mobility of network nodes as well as defend the potential attacks from enemy, a novel GT-SaRE-MANET (game theoretic based Situation-aware Robot Enhanced Mobile Ad-hoc Net- work) has been developed in this paper. (see Fig. 1). The objectives of GT-SaRE-MANET are 1) generating a unified framework effectively engaging the MANET routing with multi-robot mobility, and 2) developing a novel game theoretic situation-aware online reinforcement learning that can optimize the routing protocol as well as multi-robot path planning under uncertain and vulnerable environment at the tactical edge. To realize those objectives, a novel set of two-player (i.e. attacker and defender) game theoretic mission oriented metrics has been designed firstly that can simultaneously describe the effectiveness of MANET routing protocol, practical multi-robot mobility, and effect from vulnerable environment. Then, an optimal design problem of MANET routing and multi-robot path planning are formulated under uncertain and vulnerable environment. A game theoretic situation-aware online reinforcement learning has been developed that can learn the optimal design of MANET routing and multi-robot moving path online at tactical edge even under harsh environment. Both numerical simulation and experimental tests results have been provided to demonstrate the effectiveness of the proposed GT-SaRE-MANET protocol. Compared with Optimized Link State Routing (OLSR) [12] and emerging social-aware protocols (e.g. BubbleRap, Simbet), the proposed GT-SaRE-MANET protocol can effectively utilize the multi-robot mobility and take the affects from real-time enemy attacks into consideration. Therefore, GT-SaRE-MANET can not only significantly reduce the message delivery cost and delay but also increase the message delivery ratio even at uncertain and vulnerable tactical edge.

Beyond the battle field scenario, The GT-SaRE-MANET protocol is also expected to be used in the category of mobile networks such as land mobile networks [13], public safety network [14], interstellar networks [15], and vehicle networks [16]. Considering public safety network as an example, due to rapidly growth of population, more people has moved from urban close to suburban such as the forests where are flammable, and more new houses have been built on the fire alarm line. While a serious wildfire occurring such as paradise wildfire at California 2018 [17], it is very difficult to maintain the high-quality communication network at those areas. The spread trend and distribution of fires are even more difficult for residuals in disaster centers to predict. By joining the concept of robot-enhanced collaborative disaster relief, the autonomous robot groups can form the opportunity mobile network through mutual cooperation autonomously. Using this robot-enhanced network, the distribution and trend of the fire will be identified effectively, furthermore the critical public safety information in disaster area can be transmitted to rescue team timely for efficiently reducing the fire risk.

The rest paper is organized as follows. Section 2 presents GT-SaRE-MANET protocol and develops the game theoretic situation-aware reinforcement learning technique for both unicast and multicast scenarios. Section 3 provides the simulation settings. Section 4 demonstrates the numerial simulation results and compares GT-SaRE-MANET with OLSR, BubbleRap and Simbet, then extends the simulations to experimental tests. Section 5 concludes the paper.

2. Game Theoretic Reinforcement Learning Based Intelligent GT-SARE-MANET Routing

In this section, the development of GT-SaRE-MANET (game theoretic Situation-aware Robot Enhanced Mobile Ad-Hoc Network) routing protocol is given. It is based on the opportunistic mobile networking schemes, where a node receives packets, stores them in their buffers, carries them while moving, and forwards them to other nodes when they encounter each other.

After introducing a novel set of game theoretic mission oriented metrics, an optimal routing and multi-robot path planning design problem can be formulated. Due to the uncertainty and vulnerability of harsh environment,e.g. tactical edge, formulated optimal design can not be obtained in real-time. Therefore, game theoretic situation-aware online reinforcement learning is developed to learn the optimal routing and multi-robot moving plan that cannot only reduce message delivery delay and cost, increase message delivery ratio, and also better defend the potential worst attacks from enemy. Moreover, both unicast and multicast scenarios have been considered.

2.1 Game Theoretic Mission Oriented Metrics

The main mission of GT-SaRE-MANET routing protocol is to optimize message delivery performance as well as avoiding the potential attack by effectively using the mobility of multi-robots. To develop the effective routing for GT-SaRE-MANET, an optimal and resilient design problem needs to be formulated firstly. Considering unicast as a mission, the optimal routing design can be formulated as Markov Decision Process (MDP). However, different from conventional MANET, developed GT-SaRE-MANET needs to consider optimal design for communication network, multi-robot path planning under uncertain enemy attacks. Therefore, the mission oriented state and action space need to be redefined for GT-SaRE-MANET as

Attack observation space: $\boldsymbol{O}_i(t) = [o_{i,1}(t), ..., o_{i,l}(t)]$
State space: $\boldsymbol{S}_i(t) = [\boldsymbol{S}_{i,net}(t), \boldsymbol{S}_{i,robot}(t)], i = 1, 2, ..., N.$
Action space: $\boldsymbol{A}_i(t) = [\boldsymbol{a}_{i,net}(t), \boldsymbol{a}_{i,robot}(t)], i = 1, 2, ..., N.$

where $o_{i,1}(t)$ is the observed attacks by robot i at time t. Moreover, $\boldsymbol{S}_{i,net}(t)$,$\boldsymbol{S}_i$, $robot(t)$,$\boldsymbol{a}_{i,net}(t)$,$\boldsymbol{a}_{i,robot}(t)$ denote the robot i's mission oriented states and actions sub spaces from network routing and robot path planning aspects respectively. Specifically, $\boldsymbol{S}_{i,net}(t)$ includes mission oriented states as “keep message”, “not keep message”, “never had message”, “had message before” and “local routing table”. $\boldsymbol{S}_{i,robot}(t)$ includes “position”, “velocity”, and “acceleration” of robot i. In addition, $\boldsymbol{a}_{i,net}(t)$ includes mission oriented actions as “carrying message”, “forwarding message” and “sharing local routing table with contacts”. $\boldsymbol{a}_{i,robot}(t)$ includes “moving directions”, “moving velocity” and “moving acceleration” of robot i.

Also due to the uncertainty and vulnerability of harsh environment, e.g. tactical edge, multi-robot are placed in a distributed manner without knowing the full knowledge of network topology and possible attacking information. Therefore, directly utilizing mobility of multi-robots to enhance MANET quality is very difficult and unrealistic. To better reap the advantages from multi-robot mobility even under uncertain and vulnerable environment, a set of distributed game theoretic mission oriented metrics is developed. Those metrics in each robot are the critical performance indices that can help distributed robots better planning their mobility and routing protocol to accomplish the mission as well as defend enemy attacks effectively.

1) *Movement activity index (MA)*: degree of robot i's activity measured at time t, i.e.

$$MA_i = \int_0^t \left(v_i t + u_i t^2\right) d\tau \tag{1}$$

where v_i, u_i are the velocity and accelerator of robot i that follow the mobile robot dynamics, i.e.

$$\begin{cases} \dot{p}_i = v_i \\ \dot{v}_i = u_i \end{cases}, i = 1, 2, ..., N \tag{2}$$

where p_i and N represents the position of robot i and total number of robots in the network respectively. It is important to note that more active robot could have higher chance to find and transmit message to destination.

2) *Frequency of encounters (FE) [11]*: percentage of time (measured at time t) robot i and its neighbors are with in their communication range r_c, i.e.

$$FE_i(t) = \frac{1}{t} \int_0^t num(N_i(\tau)) d\tau \tag{3}$$

where num is the statistics of neighbors in current neighbor set

$$N_i = \{\ j \in (1, ..., N) : ||p_j(t) - p_i(t)|| \leq r_c\} \tag{4}$$

3) *Mission completion success probability (MCSP)*: current probability (measured at time t) that robot i can be used to carry and transmit message to receiver successfully which is defined as

$$MCSP_i(t) = \frac{1}{sizeof(\boldsymbol{A}_i(t))} \tag{5}$$

where $\boldsymbol{A}_i(t)$ is defined as current mission oriented actions set at robot i including mobility pattern and routing actions. The $\boldsymbol{A}_i(t)$ is updated along with time as

$$\boldsymbol{A}_i(t) = \boldsymbol{A}_i(t-1) \cap \overline{\boldsymbol{A}}_{i,hist}(t) \cap \boldsymbol{A}_{i,neighbors}(t) \tag{6}$$

with $\boldsymbol{A}_{i,neighbors}(t)$ being the mission oriented actions from robot i's neighbors, i.e. robots that contact robot i at time t. $\overline{\boldsymbol{A}}_{i,hist}(t)$ is the complementary set of robot i's historical mission oriented actions sets. When the robot operated one action and has not benefited the mission completion, that action will be stored in the historical mission oriented actions sets, i.e. $\boldsymbol{A}_{i,hist}$. It is important to note that robot will have more chance to complete the mission when it is very clear what to do, i.e. the size of mission oriented actions set is getting smaller.

4) *Potential Attacking Area Estimator (PAaE)*: The potential attacking area estimated by robot i (measured at time t), which can be defined as

$$PAaE_i(t) = g(\boldsymbol{P}_i(t), \boldsymbol{O}_i(t)) \tag{7}$$

with $\boldsymbol{O}_i(t)$, $\boldsymbol{P}_i(t)$ being defined as moving actions and positions of attackers estimated at robot i. The $\boldsymbol{P}_i(t)$ has been updated along with time as

$$\boldsymbol{P}_i(t) = \boldsymbol{P}_i(t-1) \cup \boldsymbol{P}_{i,hist}(t) \cup \boldsymbol{P}_{i,neighbors}(t) \tag{8}$$

where $\boldsymbol{P}_{i,neighbors}(t)$ is the attackers' position estimated by robot i's neighbors at time t. $\boldsymbol{P}_{i,hist}(t)$ is robot i's historical information about estimated attackers position. If robot encountered the attackers, that practical attackers' position information will be used to adjust estimation scheme. Based on the estimated attackers' position and moving action, the potential attacking area at time t can be estimated.

2.2 Game Theoretic Reinforcement Learning Based Intelligent GT-SaRE-MANET Routing

The reward function at robot i can be defined as

$$Reward\colon\ r_i(t) = f(MA_i(t), FE_i(t), MCSP_i(t), PAaE_i(t))$$

with $f(*)$ being the reward evaluation function. According to the defined game theoretic mission oriented metrics, $MA_i(t), FE_i(t), MCSP_i(t), PAaE_i(t)$ depends on the current state, action of robot i and the attackers' information observed by robot i. Then, the optimal value function corresponding to a game theoretic mission oriented policy solved as

$$V^*(t) = \max_{\pi\in PD(\boldsymbol{A})}\ \min_{o\in\boldsymbol{O}} \sum_{i=1}^{N}\sum_{\tau=t}^{T_F}\beta^\tau E(r_i(t)|\pi_i, s_i) \tag{9}$$

$$\boldsymbol{\pi}_i^*(\boldsymbol{s}_i, t) \leftarrow arg\ \max_{\pi_i\in PD(\boldsymbol{A}_i)}\ \min_{\boldsymbol{o}_i\in\boldsymbol{O}_i} \sum_{i=1}^{N} V_i^*(t) \tag{10}$$

where $\boldsymbol{o_i}$ is the observed attacker information, $r_i(t)$ is the reward at time t, $\beta^\tau \in [0, 1)$ is the discount factor at time τ. $V^*(t)$ represents the optimal value under policy π. While policy π is used to represent the probability of taking action a in state s at time t, also can be explained as the policy for action selection, so π is a probabilistic outcome, a collection of probability distributions over the available actions, $\pi \in PD(\boldsymbol{A})$.

However, it is very difficult and even impossible to attain optimal value function and policy directly due to two major challenges, i.e. 1) the harsh environment with uncertain attacks from enemy, 2) total value function, i.e. $V(t)$, cannot be obtained since multi-robots are placed in a distributed manner with limited information exchange. To overcome above challenges, an online distributed game theoretic Q-learning has developed that can learn the optimal routing, multi-robot mobility by considering worst enemy attack at harsh environment e.g. tactical edge.

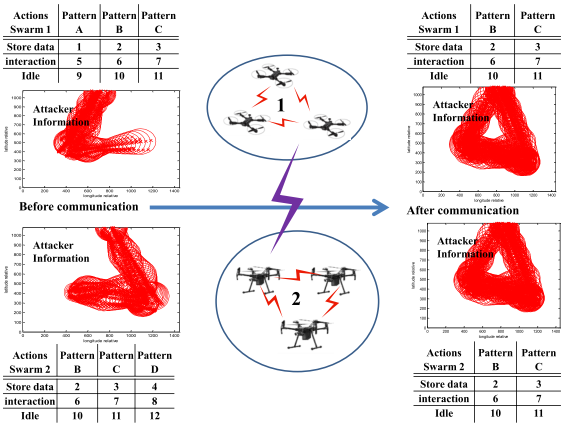
Inspired from recent Q-learning [18] and Equilibrium Q learners literatures [19]–[24], a game theoretic mission oriented optimal Q-function, can be defined as

$$Q^*(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{o}) = \sum_{i=1}^{N} Q_i^*(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i) = \sum_{i=1}^{N} V_i^*(t) \tag{11}$$

Then, the optimal game theoretic mission oriented policy, i.e. routing and robot's path planning, can be obtained as

$$\boldsymbol{\pi}_i^*(\boldsymbol{s}_i, t) \leftarrow arg\ \max_{\pi_i\in PD(\boldsymbol{A}_i)}\ \min_{\boldsymbol{o}_i\in\boldsymbol{O}_i} \sum_{i=1}^{N} Q_i^*(\boldsymbol{s}_i, \boldsymbol{a}, \boldsymbol{o}, t) \tag{12}$$

Since optimal Q-function is very difficult to obtain, we need to estimate it through Q-learning technique. Specifically, the optimal Q-function will be learned in a distributed manner as



Actions Swarm 1	Pattern A	Pattern B	Pattern C
Store data	1	2	3
interaction	5	6	7
Idle	9	10	11

Actions Swarm 2	Pattern B	Pattern C	Pattern D
Store data	2	3	4
interaction	6	7	8
Idle	10	11	12

Actions Swarm 1	Pattern B	Pattern C
Store data	2	3
interaction	6	7
Idle	10	11

Actions Swarm 2	Pattern B	Pattern C
Store data	2	3
interaction	6	7
Idle	10	11

Fig. 2 Information exchange during contact.

$$\begin{aligned} Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t+1) = Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t) + \alpha\{\, r_i(t) + \\ \gamma[V_i(\boldsymbol{s}_i^{'}, t) + V_{-i}(\boldsymbol{s}_i^{'}, t)] - Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t)\,\} \end{aligned} \tag{13}$$

$$\begin{aligned} V_i(\boldsymbol{s}_i^{'}, t) = \max_{\pi_i\in PD(\boldsymbol{A}_i)}\ \min_{\boldsymbol{o}_i\in\boldsymbol{O}_i} \sum_{\boldsymbol{a}_i\in\boldsymbol{A}_i} \{\, \pi_i\left(\boldsymbol{s}_i^{'}, \boldsymbol{a}_i^{'}, t\right) \\ \times Q_i(\boldsymbol{s}_i^{'}, \boldsymbol{a}_i^{'}, \boldsymbol{o}_i^{'}, t)\,\} \end{aligned} \tag{14}$$

$$\begin{aligned} V_{-i}^{'}(\boldsymbol{s}_i^{'}, t) = \max_{\pi_{-i}\in PD(\boldsymbol{A}_i)}\ \min_{\boldsymbol{o}_i\in\boldsymbol{O}_i} \sum_{\boldsymbol{a}_i\in\boldsymbol{A}_i} \{ \\ \pi_{-i}\left(\boldsymbol{s}_i^{'}, \boldsymbol{a}_i^{'}, \boldsymbol{a}_{-i}^{'}, t\right) \times Q_{-i}(\boldsymbol{s}_i^{'}, \boldsymbol{a}_i^{'}, \boldsymbol{a}_{-i}^{'}, \boldsymbol{o}_i^{'}, \boldsymbol{o}_{-i}^{'}, t)\,\} \end{aligned} \tag{15}$$

where α is a learning rate, γ is a discounting factor, $a^{'}, o^{'}$ denote mission oriented action and attacker's moving action at next time that can be selected from in the action space $\mathbf{A}$ and attacker observation space $\boldsymbol{O}$. And $s^{'}$ is the mission oriented state at next time that belongs to the state space $\mathbf{S}$. $V_i(\boldsymbol{s}_i^{'}, t)$ is the value of a mission oriented policy for robot i at time t. Moreover, $Q_{-i}(\boldsymbol{s}_i^{'}, \boldsymbol{a}_i^{'}, \boldsymbol{a}_{-i}^{'}, \boldsymbol{o}_i^{'}, \boldsymbol{o}_{-i}^{'}, t)$ denotes the estimate Q-function from the neighbors of robot i at time t, also the $V_{-i}^{'}(\boldsymbol{s}_i^{'}, t)$. Although multi-robots are deployed in a distributed manner, they could contact with each other when they move close within a certain region. Once different robots had the contact, they will share their current learnt Q-function, current and historical mission oriented states and actions, also observed attacker informations. Through using those information, distributed robot can better solve its own optimal policy including routing and mobility. A detailed example is given in Fig. 2. for better explaining this scenario.

Next, the estimated optimal mission oriented policy π for agent i (i.e. Routing and multi-robot path planning for GT-SaRE-MANET) can be developed as

$$\begin{aligned} \boldsymbol{\pi}_i(\boldsymbol{s}_i, t) \leftarrow arg\ \max_{\pi_i\in PD(\boldsymbol{A}_i)}\ \min_{\boldsymbol{o}_i\in\boldsymbol{O}_i} \sum_{\boldsymbol{a}_i\in\boldsymbol{A}_i} \{ \\ \boldsymbol{\pi}_i(\boldsymbol{s}_i, \boldsymbol{a}_i, t) \times Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t)\,\} \end{aligned} \tag{16}$$

2.3 Extension to Multicast Case

Multicast broadcasting is more critical and challenging especially under uncertain and vulnerable environment, e.g. tactical edge. To obtain learning based intelligent GT-SaRE-MANET in multicase case, we could extend from GT-SaRE-MANET in unicast case that developed in Sections II.A and II.B. Specifically, there are two major differences between unicast GT-SaRE-MANET and multicast GT-SaRE-MANET. First, in multicast case, there are multiple destinations for data delivery, i.e. $D_1, D_2, D_3 ..., D_M$. Therefore, the develop GT-SaRE-MANET routing in multicast case needs to maximize the possibility to find all destinations and accomplish successful message delivery. To realize this, we consider that the robot can still carry the same data message even this robot also decide to forward the message to the other robots contacted at current t. Second, the number of destinations who have not received messages needs to be updated dynamically. If one destination D_k has already received the message, the destination should stop requesting robots to forward message to it. Hence, when robot i move close to one destination, the destination will notify the robot i that "no need to come here, robot i has already sent the data", if this destination has already received the message from another robot j. Then, robot i will remove the relevant mission oriented actions. Namely, $\boldsymbol{a}_i(t)$ will be updated from

$$\boldsymbol{A}_i(t) = \boldsymbol{A}_i(t-1) \cap \overline{\boldsymbol{A}}_{i,hist}(t) \cap \boldsymbol{A}_{i,neighbors}(t) \cap \overline{\boldsymbol{A}}_{j,final}(t)$$

It is important to note that the online learning process will be terminated only if there are M robots arrived their final states, i.e. all M destinations received message from those M robots.

$$\boldsymbol{a}_i(t) = \underset{\boldsymbol{a}_i \in \boldsymbol{A}_i}{argmax} Q_i(\boldsymbol{s}_i, \boldsymbol{a}, t) \tag{17}$$

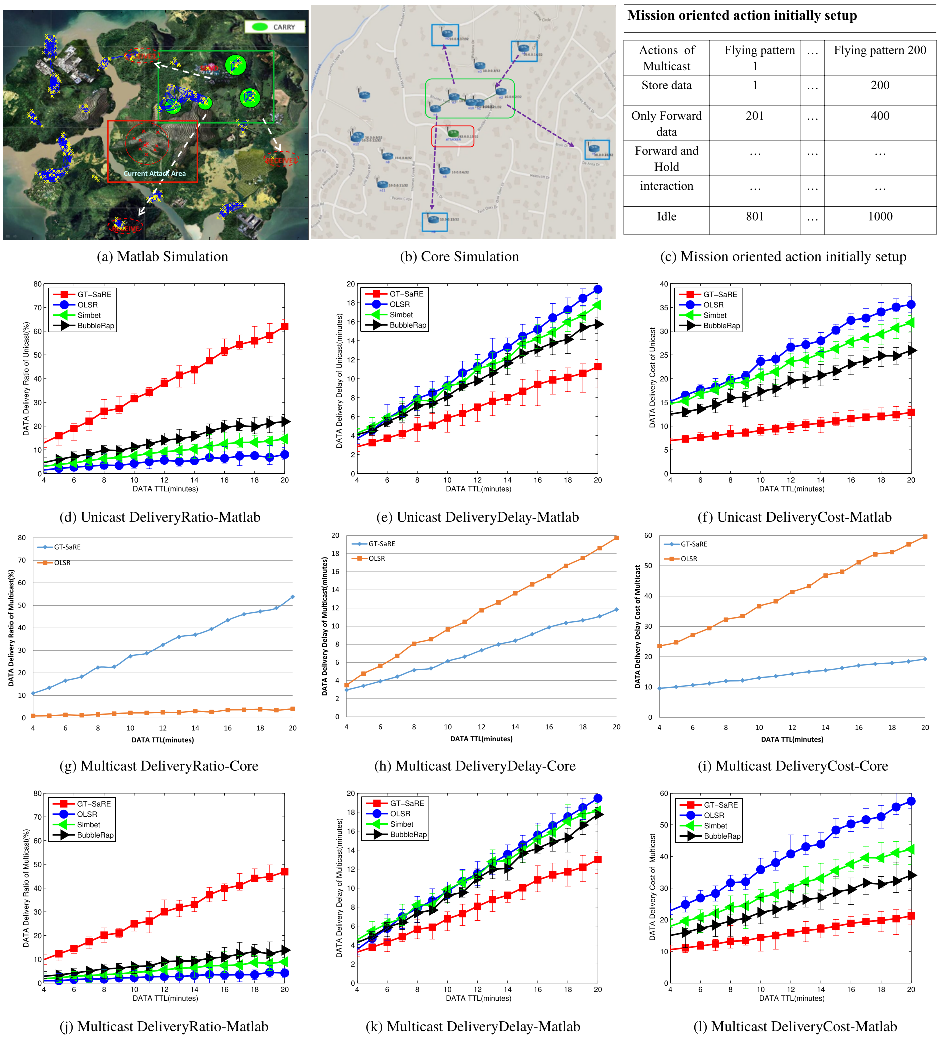
To better demonstrate the developed distributed Q-learning based intelligent design, a detailed GT-SaRE-MANET algorithm is given in the in Algorithm table 1. First, each robot has their own Q-function. Second, Each robot will obtain current reward by calculating the reward function. There is a Potential Attacking Area Estimator (PAaE) in the reward function (The potential attacking area estimated by robot i(measured at time t)) which help give a negative reward when encounter attacker. The robot i_{th} state $\boldsymbol{S}$ also includes position information. So when current robot with a specific action $\boldsymbol{a}$ at current states get a negative reward, the negative reward in its Q-function will help the robot dodge attacker. Third, Robots could share information with their neighbor since robots in the neighborhood are within communication range that can support the information exchange. Also, using the identification index, robots can effectively disguise the teammates from enemy. Using the updated information, the robots will recalculate their own Q-function. It will help each robot have an overall overview condition of attacker.

Algorithm 1 Distributed Learning Based GT-SaRE-MANET

```
Initialize the mission oriented state space $S_i$,action space $A_i$, $Oi$, policy $\boldsymbol{\pi}_i(\boldsymbol{s}_i, t)$, attacker's position information $P_i(t)$, the Q-learning environment
for $t = 1 : T_F$ do
  for $i = 1 : N$ do
    Attain current Q-function $Q_i(t)$. And find current mission oriented state $(\boldsymbol{s}_i(t))$
    estimated attackers position information update $\boldsymbol{P}_i(t) = \boldsymbol{P}_i(t-1) \cup \boldsymbol{P}_{i,hist}(t) \cup \boldsymbol{P}_{i,neighbors}(t)$
    Obtain current mission oriented action $\boldsymbol{a}_i \leftarrow \boldsymbol{\pi}_i(\boldsymbol{s}_i, t)$ Mixed with $\boldsymbol{A}_i(t) = \boldsymbol{A}_i(t-1) \cap \overline{\boldsymbol{A}}_{i,hist}(t) \cap \boldsymbol{A}_{i,neighbors}(t)$
    Estimate optimal Q-function:
    Compute the reward
      $r_i(t) = f(MA_i(t), FE_i(t), MCSP_i(t), PAaE_i(t))$
    Compute
      $Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t+1) = Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t) + \alpha \{ r_i(t) + \gamma[V_i(\boldsymbol{s}_i', t) + V_{-i}(\boldsymbol{s}_i', t)] - Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t) \}$
      $V_i(\boldsymbol{s}_i', t) = \underset{\pi_i \in PD(\boldsymbol{A}_i)}{max} \underset{\boldsymbol{o}_i \in \boldsymbol{O}_i}{min} \sum_{\boldsymbol{a}_i \in \boldsymbol{A}_i} \{ \pi_i \left(\boldsymbol{s}_i', \boldsymbol{a}_i', t\right) \times Q_i(\boldsymbol{s}_i', \boldsymbol{a}_i', \boldsymbol{o}_i', t) \}$
      $V_{-i}'(\boldsymbol{s}_i', t) = \underset{\pi_{-i} \in PD(\boldsymbol{A}_i)}{max} \underset{\boldsymbol{o}_i \in \boldsymbol{O}_i}{min} \sum_{\boldsymbol{a}_i \in \boldsymbol{A}_i} \{ \pi_{-i} \left(\boldsymbol{s}_i', \boldsymbol{a}_i', \boldsymbol{a}_{-i}', t\right) \times Q_{-i}(\boldsymbol{s}_i', \boldsymbol{a}_i', \boldsymbol{a}_{-i}', \boldsymbol{o}_i', \boldsymbol{o}_{-i}', t) \}$
    Estimate Optimal $\boldsymbol{\pi}_i(\boldsymbol{s}_i, t)$:
      $\boldsymbol{\pi}_i(\boldsymbol{s}_i, t) \leftarrow arg \underset{\pi_i \in PD(\boldsymbol{A}_i)}{max} \underset{\boldsymbol{o}_i \in \boldsymbol{O}_i}{min} \sum_{\boldsymbol{a}_i \in \boldsymbol{A}_i} \{ \pi_i(\boldsymbol{s}_i, \boldsymbol{a}_i, t) \times Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{o}_i, t) \}$
  end
  Update Q-function,i.e. $Q_i(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{o}, t)$ and Update Q-learning environments
  Stop learning if one robot arrives final state, i.e. successfully deliver message to destination
end
```

In the unicast GT-SARE-MANET simulation, the mission oriented action for each UAVs group has been defined as $\mathbf{a}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}]$, with $\mathbf{a}_{i,1}$ being the mission oriented action vector that represents the flying patterns. Initially, 200 flying patterns has been generated, i.e. $\mathbf{a}_{i,1} = [a_{i,11} a_{i,12} ... a_{i,1\,200}]$. Next, $\mathbf{a}_2$ is the action vector about four routing choices ["store data"; "forward data"; "interaction"; "idle"]. Therefore, the mission oriented action table for each UAVs group initially contains 4*200 elements. In multicast GT-SaRE-MANET simulation, the mission oriented action has been defined as $\mathbf{a} = [\mathbf{a}_1, \mathbf{a}_2]$, with $\mathbf{a}_1$ being the mission oriented action vector about flying patterns. However, in the multicast case, multi-UAV may still hold the same necessary data after forwarding. Hence, the $\mathbf{a}_2$ needs to be defined as ["store data"; "only forward data"; "forward and hold data"; "interaction";"idle"]. Therefore, the mission oriented action table for each UAVs group initially contains 5*200 elements.

The destination positions of message transmitter and receiver have been initialized randomly for both unicast and



(a) Matlab Simulation

(b) Core Simulation

Mission oriented action initially setup

Actions of Multicast	Flying pattern 1	…	Flying pattern 200
Store data	1	…	200
Only Forward data	201	…	400
Forward and Hold	…	…	…
interaction	…	…	…
Idle	801	…	1000

(c) Mission oriented action initially setup

(d) Unicast DeliveryRatio-Matlab

(e) Unicast DeliveryDelay-Matlab

(f) Unicast DeliveryCost-Matlab

(g) Multicast DeliveryRatio-Core

(h) Multicast DeliveryDelay-Core

(i) Multicast DeliveryCost-Core

(j) Multicast DeliveryRatio-Matlab

(k) Multicast DeliveryDelay-Matlab

(l) Multicast DeliveryCost-Matlab

Fig. 3 Performance evaluation.

multicast cases. After 1000 times indices training, three critical parameters can be computed and used to evaluate the performance of different routing protocols, i.e.,

- **Data delivery ratio**: ratio of data message successfully received to the total sent.
- **Data delivery delay**: the average time for the destinations to receive the data message from transmitter.
- **Data delivery cost**: the number of times that a data message is forwarded before delivered.

3. Simulation Setup

To better demonstrate the effectiveness of proposed intelligent GT-SaRE-MANET routing, both matlab and Common Open Research Emulator (CORE) have been used in the computer aided simulation. Also several recent network protocols including OLSR, Simbet, Bubblerap have been used as benchmarks to compare with proposed design. Moreover, we are interested in effects from the uncertain and vulnerable environment to the network protocol. In the simulation, the network traffic was generated with random source and actions. To simulate the transient connectivity in practical battlefield, the unmanned aerial vehicle (UAV), commonly known as a flying robot is considered here. Also the Reference Group mobility model has been used. It is important to note that developed GT-SaRE-MANET has been tested at the same scenario. However, since the multi-UAV mobility and routing design all based on reinforcement learning, there is no need to set explicit mobility mode. In GT-SaRE-MANET simulation, multi-UAV have been divided into different groups. Each group has a leading UAV determining the mobility of entire group. We use the same experiment settings such that the attackers' movements is initialized with specific trajectory, and the message is generated randomly and data time-to-live (TTL) varies from 5 minutes to 20 minutes.

4. Simulation Results

In this section, the performance of developed game theoretic Q-learning based GT-SaRE-MANET routing has been eval-

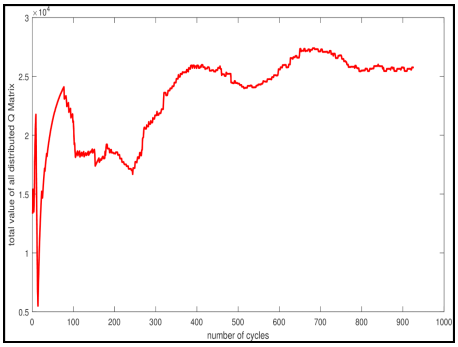


Fig. 4 Global reward evaluation for multicast case.

uated and compared with several recent routing protocols, including OLSR protocol, SimBet, BubbleRap, especially at tactical edge. In Fig. 3(d) to (f), the performance of develop learning based GT-SaRE-MANET routing for unicast case has been evaluated in matlab emulator. Compared with recent advanced network routing protocol, the developed GT-SaRE-MANET can maximize the message delivery ratio as well as minimize the message delivery delay and cost. Even under the uncertain and vulnerable harsh environment such as tactical edge, proposed GT-SaRE-MANET can provide over 60% message delivery ratio at tactical edge. It is because proposed design has intelligently utilized game theoretic based situation aware online reinforcement learning to effectively manage the routing protocol as well as multi-robot mobility even under uncertainty enemy attacks. Then, the performance of proposed GT-SaRE-MANET routing for multi-cast at tactical edge has been demonstrated in Fig. 3(g) to (l) separately via matlab and CORE emulator. They both give similar results. Although multicast case is much more complicate and difficulty than unicast case, the developed learning based GT-SaRE-MANET routing still can provide a much more efficient and reliable performance even at tactical edge, where the distributed Q function is being converged and shown in Fig. 4.

To consider more realistic uncertainty and vulnerability from tactical environment, a battlefield communication scenario has been set up and used to evaluate the performance of designed scheme. In this scenario, a tactical map has been adopted and several military groups have been divided into different areas. After randomly selected transmitter and receiver within military groups, the mission UAVs will be initialized in different places and used to help on message exchange. And the attacking UAV group is set up with specific trajectory. Based on the tactical map, the network model will be generated and included in real-time simulation. Moreover, mission UAVs do not have fully knowledge of transmitter and receiver locations and actual attackers' trajectory. Therefore, the designed algorithm will advise distributed mission UAVs to search transmitter and receiver, avoid the potential attack and then carry and forward the message from source to destination.

Both unicast and multicast cases have been investigated. As shown in Fig. 3(a) (b), the best group of UAV (i.e. green frame) will carry and forward message to destination successfully. In this realistic battlefield scenario, the developed algorithm can provide better performance than recent advanced routing protocol, i.e. lower delivery delay and cost, and higher deliver ratio. To validate the practical effectiveness of the proposed learning based GT-SaRE-MANET routing algorithm, a real-time experimental test has been conducted. We set up three DJI S1000 UAVs, two among them has been rebuilt by including 5.8 GHz wireless transceiver module and an Intel i7 processor for embedding online game theoretic reinforcement learning algorithm. The Fig. 5 show the detection radar setup to detect attackers position in real-time. The outdoor test has been run at Federal Aviation Association (FAA) certified UAS test site. During the test, message transmitter and receiver have been deployed firstly. Then, two UAVs has been deployed in the different places without known the location of transmitter, receiver and each other. Then, one UAV act as attacker flying with specific trajectory to interrupt the message exchange. In Fig. 6, the developed GT-SaRE-MANET scheme can effectively force two distributed UAVs to not only find the transmitter and receiver, and also efficiently avoid the attacking UAV and transmit the message from source to the destination even under the uncertain and vulnerable environment.

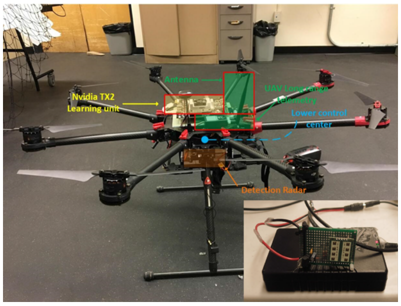


Fig. 5 DJI S1000 platform and detection radar.

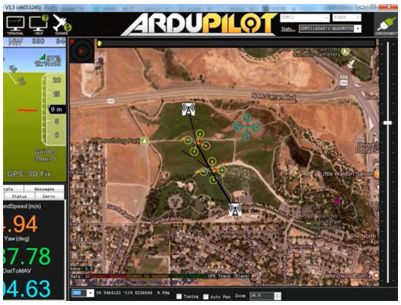


Fig. 6 Real-time ground station monitor.

5. Conclusion

In this paper, a novel multi-robot enhanced MANET has been investigated. Through integrating the advanced online game theoretic reinforcement learning technique, a series of intelligent network routing and multi-robot mobility can be obtained to not only significantly improve the MANET quality but also handle the practical effects from uncertain and vulnerable harsh environment such as the battlefield. Moreover, the effectiveness of the proposed algorithm has been validated through computer-aid simulations as well as real-time experimental tests. Both numerical simulation and experimental results demonstrated that the proposed online learning based GT-SaRE-MANET scheme can provide much better performance than the state-of-the-art designs such as OLSR, BubbleRap, Simbet, especially in uncertain and vulnerable environment of tactical edge.

References

[1] B. Yang, Y. Chen, G. Chen, and X. Jiang, "Throughput capacity study for MANETs with erasure coding and packet replication," IEICE Trans. Commun., vol.E98-B, no.8, pp.1537–1552, Aug. 2015.

[2] Y. Ochi, K. Kinoshita, H. Tode, and K. Murakami, "A design of wide area MANET by dynamic linkage with IP-based infrastructure," IEICE Trans. Commun., vol.E92-B, no.3, pp.889–897, March 2009.

[3] K. Okano, Y. Aoki, T. Ohta, and Y. Kakuda, "An inter-domain routing protocol based on autonomous clustering for heterogeneous mobile ad hoc networks," IEICE Trans. Commun., vol.E98-B, no.9, pp.1768–1776, Sept. 2015.

[4] S.U. Rehman, M.A. Khan, and T.A. Zia, "A multi-hop cross layer decision based routing for VANETs," Wirel. Netw., vol.21, no.5, pp.1647–1660, 2015.

[5] K. Fall, "A delay-tolerant network architecture for challenged internets," Proc. 2003 conference on Applications, technologies, architectures, and protocols for computer communications, pp.27–34, ACM, 2003.

[6] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: Data forwarding in disconnected mobile ad hoc networks," IEEE Commun. Mag., vol.44, no.11, pp.134–141, 2006.

[7] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," Duke Technical Report, CS-2000-06, 2000.

[8] E. Daly and M. Haahr, "Social network analysis for routing in disconnected delay tolerant MANETs," Proc. 8th ACM Intl. Sym. on Mobile Ad Hoc Networking and Computing, Montreal, Canada, Sept. 2007.

[9] P. Hui, J. Crowcroft, and E. Yoneki, "BUBBLE rap: Social-based forwarding in delay tolerant networks," MobiHoc, pp.241–250, 2008.

[10] E. Bulut and B.K. Szymanski, "Exploiting friendship relations for efficient routing in mobile social networks," IEEE Trans. Parallel Distrib. Syst., vol.23, no.12, pp.2254–2265, Dec. 2012.

[11] Y.E. Sagduyu, Y. Shi, T. Erpek, S. Soltani, S.J. Mackey, D.H. Cansever, M.P. Patel, B.F. Panettieri, B.K. Szymanski, and G. Cao, "Multilayer MANET routing with social-cognitive learning," MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM), pp.103–108, 2017.

[12] T. Clausen and P. Jacquet, "Optimized link state routing protocol (OLSR)," Network Working Group, RFC 3626, Oct. 2003.

[13] M. Sprinkle, "Design considerations in a modern land mobile radio system," Diss. Virginia Tech., 2003.

[14] T. Ohtsuji, K. Muraoka, H. Aminaka, D. Kanetomo, and Y. Matsunaga, "Relay selection scheme based on path throughput for device-to-device communication in public safety LTE," IEICE Trans. Commun., vol.E101-B, no.5, pp.1319–1327, May 2018.

[15] J.A. Fraire, M. Feldmann, F. Walter, E. Fantino, and S.C. Burleigh, "Networking in interstellar dimensions: Communicating with TRAPPIST-1," IEEE Trans. Aerosp. Electron. Syst., 2018.

[16] Y. Wang, K. Venugopal, A.F. Molisch, and R.W. Heath, "MmWave vehicle-to-infrastructure communication: Analysis of urban microcellular networks," IEEE Trans. Veh. Technol., vol.67, no.8, pp.7086–7100, 2018.

[17] N.L. Caldararo, "Paradise wild fire: You cannot blame it all on climate change, home survival, defensive space and self-fulfilling prophesy," Home Survival, Defensive Space and Self-fulfilling Prophesy, Nov. 2018.

[18] P. Dayan and C. Watkins, "Q-learning," Mach. Learn., vol.8, no.3/4, pp.279–292, 1992.

[19] J. Foerster, Y.M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," Advances in Neural Information Processing Systems, 2016.

[20] M. Littman and J. Boyan, "A distributed reinforcement learning scheme for network routing," Proc. First International Workshop on Applications of Neural Networks to Telecommunications, pp.45–51, 1993.

[21] A. Elwhishi, P.-H. Ho, K. Naik, and B. Shihada, "ARBR: Adaptive reinforcement-based routing for DTN," Proc. WIMOB IEEE 6th international conference wireless and mobile computes, networks and communications, pp.376–385, 2010.

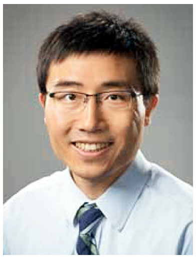
[22] V.G. Rolla and M. Curado, "A reinforcement learning-based routing for delay tolerant networks," Eng. Appl. Artif. Intell., vol.26, no.10, pp.2243–2250, 2013.

[23] J. Hu and M.P. Wellman, "Nash Q-learning for general-sum stochastic games," J. Mach. Learn. Res., vol.4, pp.1039–1069, 2003.

[24] M.L. Littman, "Friend-or-foe Q-learning in general-sum games," Eighteenth International Conference on Machine Learning, pp.322–328, Williams College, MA, 2001.

Ming Feng received his Bachelor Degree in Communication Engineering from Taiyuan University of Technology in 2011 and received M.S. degrees in Electrical Engineering from Stevens Institute of Technology in 2016. He currently worked as PHD research assistant with Dr. Hao Xu in Electrical and Biomedical Engineering at University of Nevada, Reno.

Lijun Qian is a Regents Professor and holds the AT&T Endowment in the Department of Electrical and Computer Engineering at Prairie View A&M University (PVAMU). Before joining PVAMU, he was a MTS in Bell-Labs at Murray Hill, NJ. His research interests are in the area of big data processing, wireless communications and mobile networks, network security and intrusion detection, and computational and systems biology.



Hao Xu received his Master's degree in Electrical Engineering from Southeast University in 2009, and his Ph.D. degree from the Missouri University of Science and Technology (formerly, the University of Missouri-Rolla), Rolla in 2012. He currently holds an assistant professor position with the Department of Electrical and Biomedical Engineering at University of Nevada, Reno.