PACKET FORWARDING IN MULTI-DOMAIN WIRELESS SENSOR NETWORKS USING GAME THEORETIC MULTI-AGENT REINFORCEMENT LEARNING

Sajee Singsanga

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Telecommunication Engineering

Suranaree University of Technology

Academic Year 2016

การส่งต่อแพกเกตในเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมน โดยใช้วิธีการเรียนรู้แบบมัลติเอเจนท์รีอินฟอร์สเมนท์และทฤษฎีเกมส์

นางสาวศจี สิงห์สง่า

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมโทรคมนาคม
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2559

PACKET FORWARDING IN MULTI-DOMAIN WIRELESS SENSOR NETWORKS USING GAME THEORETIC MULTI-AGENT REINFORCEMENT LEARNING

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Thesis Examining Committee

(Assoc. Prof. Dr. Monthippa Uthansakul)
Chairperson

(Asst. Prof. Dr. Wipawee Hattagam)
Member (Thesis Advisor)

(Assoc. Prof. Dr. Peerapong Uthansakul)
Member

(Assoc. Prof. Dr. Chaodit Aswakul)
Member

(Asst. Prof. Dr. Kitsuchart Pasupa)
Member

(Prof. Dr. Santi Maensiri)
Acting Vice Rector for Academic Affairs and Internationalization

(Assoc. Prof. Flt. Lt. Dr. Kontorn Chamniprasart)
Dean of Institute of Engineering

ศจี สิงห์สง่า : การส่งต่อแพกเกตในเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมนโดยใช้วิธีการเรียนรู้แบบมัลติเอเจนท์รีอินฟอร์สเมนท์และทฤษฎีเกมส์ (PACKET FORWARDING IN MULTI-DOMAIN WIRELESS SENSOR NETWORKS USING GAME THEORETIC MULTI-AGENT REINFORCEMENT LEARNING) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.วิภาวี หัตถกรรม, 164 หน้า

ปัจจุบันนี้จำนวนการใช้งานเครือข่ายเซนเซอร์ไร้สายได้เพิ่มสูงมากขึ้น เครือข่ายเซนเซอร์ไร้สายถูกนำไปใช้ในหลากหลายแอปพลิเคชั่น จึงเป็นไปได้ว่าในบริเวณพื้นที่หนึ่งๆ จะมีเครือข่ายเครือข่ายเซนเซอร์ไร้สายหลายเครือข่ายถูกใช้งานภายในบริเวณพื้นที่เดียวกันซึ่งควบคุมโดยผู้ดูแลระบบที่ต่างกัน เครือข่ายประเภทนี้เรียกว่า เครือข่ายเซนเซอร์ไร้สายแบบมัลติโดเมน อย่างไรก็ตาม เครือข่ายเหล่านี้มีแหล่งพลังงานที่จำกัด ในสถานการณ์เช่นนี้ การใช้ทรัพยากรร่วมกันระหว่างเซนเซอร์โหนดที่อยู่ต่างโดเมนอาจยืดอายุการใช้งานเครือข่ายและสร้างความน่าเชื่อถือให้กับเครือข่ายในเทอมของอัตราการส่งแพกเกตสำเร็จได้ อย่างไรก็ตาม ด้วยพฤติกรรมที่เห็นแก่ตัวของเซนเซอร์โหนดในการสงวนพลังงานที่มีอยู่อย่างจำกัด อาจไม่เอื้อให้เกิดความร่วมมือดังกล่าว ดังนั้นวิทยานิพนธ์ฉบับนี้ จึงมีวัตถุประสงค์ 1) เพื่อระบุปัจจัยที่ส่งผลกระทบต่อความร่วมมือระหว่างเครือข่ายและผลประโยชน์ที่ได้รับอย่างเท่าเทียมกันในเครือข่ายเซนเซอร์ไร้สายแบบมัลติโดเมน; 2) เพื่อประยุกต์ในทฤษฎีเกมความไม่ร่วมมือในการจัดสรรเส้นทางการส่งแพกเกตระหว่างเครือข่ายแบบกระจายสำหรับเครือข่ายเซนเซอร์ไร้สายแบบมัลติโดเมนที่มีสถานีฐานร่วมกันและสถานีฐานแยกกัน; 3) เพื่อนำเสนออัลกอริทึ่มค้นหาเส้นทางที่ได้มาซึ่งกลยุทย์ร่วมที่ดีที่สุดในการส่งต่อแพกเกตในเครือข่ายเซนเซอร์ไร้สายแบบมัลติโดเมนแบบกระจายที่พิจารณาพฤติกรรมที่ไม่ร่วมมือของเซนเซอร์โหนดด้วยการใช้กระบวนการเรียนรู้แบบรีอินฟอร์เมนท์และทฤษฎีเกม

งานวิจัยนี้ มีองค์ความรู้หลักหกประการ องค์ความรู้ประการการแรกคือ การกำหนดปัจจัยที่มีผลกระทบต่อความร่วมมือระหว่างเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมน เพื่อให้ทุกเครือข่ายได้รับผลประโยชน์ร่วมกันสูงสุด องค์ความรู้ประการที่สอง คือการออกแบบตารางผลตอบแทนสำหรับเกมการส่งต่อแพกเกตที่ประกอบด้วยผู้เล่นที่ไม่ร่วมมือกันสำหรับเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมนที่มีการจัดการแบบกระจาย องค์ความรู้ประการที่สาม คือ การนำเสนอกระบวนการค้นหาเส้นทางในเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมนที่ไม่ร่วมมือด้วยการใช้ทฤษฎีเกมความไม่ร่วมมือ องค์ความรู้ประการที่สี่ คือ การออกแบบฟังก์ชั่นคุณลักษณะที่เหมาะสมต่อวิธีแนชคิวที่มีสถานะแบบต่อเนื่อง องค์ความรู้ประการที่ห้า คือ การนำเสนอกระบวนการค้นหาเส้นทาง เพื่อให้ได้ผลประโยชน์ร่วมกันสูงสุดระหว่างเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมนที่ไม่ร่วมมือกันด้วย

การใช้วิธีแนชคิวที่มีสถานะแบบต่อเนื่อง องค์ความรู้ประการที่หก คือ การนำเสนอกระบวนการค้นหาเส้นทางที่ได้รับผลประโยชน์ร่วมกันอย่างยุติธรรมต่อทุกเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมน

ผลการทดลองชี้ให้เห็นว่า วิธีการที่นำเสนอสามารถกำหนดการส่งต่อแพกเกตสำหรับเครือข่ายเซ็นเซอร์ไร้สายแบบมัลติโดเมนได้อย่างมีประสิทธภาพ โดยวิธีการที่นำเสนอสามารถเพิ่มอายุเครือข่ายและอัตราการรับส่งแพกเกตข้อมูลได้สูงขึ้น และสามารถกำหนดเส้นทางการส่งต่อแพกเกตที่ได้รับผลประโยชน์ร่วมกันอย่างยุติธรรมต่อทุกเครือข่ายได้ นอกจากนี้ยังมีความทนทานต่อการเปลี่ยนแปลงสภาพแวดล้อมได้ดีกว่า (ได้แก่ การเปลี่ยนแปลงความหนาแน่นโหนด ค่าการสูญเสียเชิงวิถีในอากาศ จำนวนโหนดผิดพลาด รูปแบบของรูปร่างเครือข่าย และสถานะการเชื่อมต่อของเส้นทางภายในเครือข่าย)

สาขาวิชาวิศวกรรมโทรคมนาคม ลายมือชื่อนักศึกษา _____

ปีการศึกษา 2559 ลายมือชื่ออาจารย์ที่ปรึกษา _____

SAJEE SINGSANGA : PACKET FORWARDING IN MULTI-DOMAIN WIRELESS SENSOR NETWORKS USING GAME THEORETIC MULTI-AGENT REINFORCEMENT LEARNING. THESIS ADVISOR : ASST. PROF. WIPAWEE HATTAGAM, Ph.D., 164 PP.

WIRELESS SENSOR NETWORKS / MULTI-DOMAIN / NON-COOPERATIVE GAME/ LEMKE HOWSON METHOD/ DISCRETE STATE NASH Q-LEARNING (D-NASHQ)/ CONTINUOUS STATE NASH Q-LEARNING (C-NASHQ)

Wireless Sensor Networks (WSNs) have increasingly attracted much interest in a wide range of application scenarios in recent years. For certain applications, it is possible that multiple sensor networks which are controlled by different authorities can coexist independently within a region of interest. These networks may even be physically overlapping and their sensor nodes may be interleaved. Such networks are referred to as multi-domain WSNs. However, these networks usually have limitation in energy capacity. In such a situation, resource sharing and cooperation between sensor node belonging in different domain authorities may prolong network lifetime and enhance reliability on packet delivery ratio. However, selfish behaviors of sensor nodes in order to conserve their energy refuse to cooperate. Hence, the underlying objective of this thesis is to propose an adaptive routing algorithm to 1) identify the parameters that effect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve; 2) to apply non-cooperative game theory to allocate packet forwarding problem in distributed multi-domain WSNs based on common sink and separate sink scenarios; 3) to obtain routing schemes which can achieve the best mutual packet forwarding strategy in non-cooperative multi-domain

WSNs in a distributed manner using game theoretic reinforcement learning algorithm.

The main contributions of this research are six-fold. The first contribution is identification of parameters that effect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve. The second contribution is design of payoff matrix for non-cooperative packet forwarding game in distributed multi-domain WSNs. The third contribution is to propose non-cooperative game algorithm (NCG-LH) to distributed packet forwarding scheme in non-cooperative multi-domain WSNs under common sink and separate sink scenarios. The fourth contribution is design of feature function that suitable for continuous state Nash Q-learning. The fifth contribution is a proposed adaptive routing algorithms (D-NashQ and C-NashQ) and its application to packet forwarding problems in multi-domain WSNs under separate sink scenario. The sixth contribution is fair cooperative routing comparison is made between routing algorithm based on load balancing technique, non-cooperative game theory technique and game theoretic reinforcement learning technique.

The experiments show that by using the proposed algorithm which provides fair route selection, all networks can send their packets more reliably and gain longer network lifetime In addition, the proposed algorithms achieve higher robustness in changing of network condition (i.e., network density, path loss exponent, node failure, types of network topologies and connectivity).

School of Telecommunication Engineering Student's Signature _____

Academic Year 2016 Advisor's Signature _____

ACKNOWLEDGEMENT

I am grateful to all those, who by their direct or indirect involvement have helped in the completion of this thesis.

First and foremost, I would like to express my sincere thanks to my thesis advisors, Asst. Prof. Dr. Wipawee Hattagam for her invaluable help and constant encouragement throughout the course of this research. I am most grateful for her teaching and advice, not only the research methodologies but also many other methodologies in life. I would not have achieved this far and this thesis would not have been completed without all the support that I have always received from her. In addition, I am grateful for the lecturers in School of Telecommunication Engineering for their suggestion and all their help.

I would also like to thank Assoc. Prof. Dr. Monthippa Uthansakul, Assoc. Prof. Dr. Peerapong Uthansakul, Assoc. Prof. Dr. Chaodit Aswakul and Asst. Prof. Dr. Kitsuchart Pasupa for accepting to serve in my thesis examination committee.

I also grateful to the funding One Research One Gretuate (OROG) support from Suranaree University of Technology (SUT).

Finally I am most grateful to my parents and my friends both in both masters and doctoral degree courses for all their support throughout the period of this research

Sajee Singsanga

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

SYMBOLS AND ABBREVIATIONS

WSNs	=	Wireless sensor networks
QoS	=	Quality-of-service
NCG	=	Non-cooperative game
LH	=	Lemke Howson method
NE	=	Nash equilibrium
RL	=	Reinforcement learning
TD	=	Temporal difference control algorithm
MARL	=	Multi-agent reinforcement learning
NashQ	=	Nash Q-learning
MDP	=	Markov decision process
I	=	Set of agents
I	=	Agent
$-I$	=	Agent i's opponents
A	=	Set of actions
a	=	Action
a^*	=	NE action
u	=	Payoff function
t	=	Time step index
s_t	=	State of the process at time t
S	=	State space

SYMBOLS AND ABBREVIATIONS (Continued)

s	=	Current state
s'	=	Next state
$E[\cdot]$	=	Expectation operator
	=	Discount factor
~	=	Learning rate
$R(s,a,s')$	=	Expected reward at the current state s and action a with transition to next state s '
r	=	Reward
R_t	=	Expected discounted return of the agent at time t
P	=	State transition probability
®	=	Policy
$®^*$	=	Optimal policy
$E^{®}[\cdot]$	=	Expectation operator under policy ®
$V^{®}(s)$	=	Value function of a state (s) under policy ®
$V^*(s)$	=	Value function of a state (s) under optimal policy $®^*$
$Q_t^{®}(s,a)$	=	Action-value function of a given policy ® associated to state-action pair (s,a)

SYMBOLS AND ABBREVIATIONS (Continued)

$Q^*(s,a)$	=	Action-value function of a given optimal policy ®* associated to state-action pair (s,a)
E_{TX}	=	Transmission cost
E_{RX}	=	Reception cost
b	=	Size of the transmitted packet
E_{elec}	=	Expended cost in the radio electronics
±	=	Path loss exponent
£$_{amp}$	=	Energy consumed at the output transmitter antenna for transmitting one meter
N	=	Set of networks
n	=	Sensor node in the network
«	=	Packet received rate
¦	=	Cooperative energy required for cooperation
¥	=	Energy reduction obtained from changing from non-cooperative route to cooperative route
P_b	=	Bit error probability for one hop
£$_i^{nc}$	=	End-to-end energy cost along non-cooperative route for agent domain i

SYMBOLS AND ABBREVIATIONS (Continued)

$£_i^S$	=	Energy required at the source to cooperate with the other domain to forward its packet to a sink
$£_i^C$	=	Energy used by nodes in domain *i* required to help the other = domain forward their packets to a sink
$£_t(s)$	=	ε-greedy function
$K_t(s)$	=	Number of visits to state at time *t*
$K_t(a)$	=	Number of times action *a* is selected at time *t*
Φ	=	Feature function matrix
¤	=	Feature function
¨(•)	=	Indicator function
¢	=	Temporal difference error
–	=	Weight value
E_{remain}	=	Remaining battery energy for all node in the route path
E_{total}	=	Total energy consumption for packet forwarding in the route path
$E_{initial}$	=	Initial battery energy of sensor nodes in the route path
D-NashQ	=	Discrete state Nash Q-learning
C-NashQ	=	Continuous state Nash Q-learning

SYMBOLS AND ABBREVIATIONS (Continued)

D	=	Action which the agent does not forward the packet to the other network and drops all packets from other network if asked for help to forward the packets
F	=	Action which the agent forwards the packet to the other network and in turn forwards all packets if the other network asked for help to forward the packets
BER	=	Bit error rate
PRR	=	Packet received rate
OQPSK	=	Offset quadrature phase shift keying
PDR	=	Packet delivery ratio

CHAPTER I

INTRODUCTION

This chapter introduces a background problem in packet forwarding cooperation in multi-domain wireless sensor networks (WSNs) and highlights the significance of resource allocation using game theoretic reinforcement learning (GTRL) technique. It also presents the motivation for applying GTRL technique to achieve the best mutual policy for all network domains which is the main focus of this thesis.

1.1 Significance of the problem

Wireless Sensor Networks (WSNs) have increasingly attracted much interest in a wide range of application scenarios in recent years (Mattern et al., 2010; Fadel et al., 2015; Rashid et al., 2016). For certain applications, it is possible that multiple sensor networks which are controlled by different authorities can coexist independently within a region of interest. These networks may even be physically overlapping and their sensor nodes may be interleaved. Such networks are referred to as multi-domain WSNs. The networks perform different tasks and measure different data within the same area. Examples of multiple networks co-located deployments can be found in environmental monitoring with forest fire, earthquake, wildlife tracking and landslide detection sensors, and in animal monitoring with each herd belongs to a different owner.

Normally, WSNs consists of distributed autonomous sensor nodes that are often deployed in remote or hostile environments to collect and send data packet through multi-hop wireless communication to a sink in its own domain. However, these sensor nodes usually have limitation in memory size, computational capabilities and energy capacity. Since the most common energy storage device used in a sensor node is a battery which is an energy constraint, changing new battery or sensor nodes may be difficult to do in many applications. In such situations, cooperation among sensor nodes belonging to different network authorities could potentially gain certain benefits. Such benefits include alternative routing paths and reduced energy consumption, which can prolong their network lifetime and enhance reliability of packet delivery. These benefits lead to development of a new protocol with features needed in a short duration and implementation with a small cost.

However, a significant amount of energy is also lost when sensor nodes within the multi-domain WSN cooperatively process and forward the data for other networks. As energy consumption is a critical issue for such networks, reducing energy consumption and prolonging the network lifetime are important targets as shown in the following researches.

1.1.1 Cooperative routing among multi-domain WSNs

With several advantages to be gained from cooperative routing in multi-domain networks, many routing approaches have been proposed to achieve optimized energy usage in multi-domain WSNs. Most existing researches consider resource allocation problem in a *fully cooperative* situation, meaning that, the authorities have to agree on sharing or providing a common resource in order to

increase certain benefits for their networks. In (Bicakci et al., 2013 and Bicakci et al., 2010), the potential benefits of cooperation in multiple WSNs are investigated. Linear programming is employed to find an energy efficient path in order to prolong their network lifetime. However, energy efficient routing selection is not always guaranteed to prolong the network lifetime. Sensor nodes belonging to energy efficient path tend to have higher traffic load and consume more energy than other nodes. As a result, such nodes tend to die earlier. In order to avoid heavy loaded situations, Nagata et al. (2012) proposed cooperation between multi-domain WSNs by balancing the communication load. Routes with the maximum value of bottleneck were selected. By doing this, the network lifetime can be extended among multiple domains within the same geographic area. Kinoshita et al. (2016) proposed a fair cooperative routing method for heterogeneous overlapped WSNs called pool-based selecting method. An energy pool was introduced to maintain the total amount of energy consumption used in cooperative forwarding. Their simulation results showed that the proposed method was able to balance the energy consumption and prolong the network lifetime. Ref. (Jelicic et al., 2014; Singhanat et al., 2015) showed benefits of node collaboration in multi-domain WSNs under practical implementation. The results showed that cooperation with co-located sensor devices in different networks can indeed increase the network lifetime.

However, Vaz et al., (2008) and Ze et al., (2012) showed that cooperation between different networks that are deployed in the same region may not always be beneficial to every network. It is possible that some WSNs can prolong their network lifetime but shorten lifetimes of other WSNs. In Ze et al., (2012), it has been reported that the presence of only a few selfish nodes can degrade the

performance of an entire system. Thus, encouraging nodes to be cooperative and helpful in detecting selfish nodes in packet transmission is critical to ensure the proper functioning of multi-domain WSNs. Vaz et al., (2008) showed that cooperation between two authorities in co-located areas may not always be beneficial to any network, because whether or not each authority will cooperate depends on the configuration of each network. Their results showed that there are four factors which affect node cooperation, i.e. the density of the network, the data collection rate, the path loss exponent and the routing algorithm. Hence, node cooperation between different authorities in multi-domain WSNs is not straight forward.

Furthermore, multi-domain WSNs also consider fair cooperative packet forwarding for each authority in order to efficiently decide whether to cooperate with each other or not. This is of particular significance in a non-cooperative environment in order to provide fairness and benefits to all co-located networks.

1.1.2 Fair routing in multi-domain WSNs

Many researches try to find a routing algorithm which can rationally decide to select the best routing policy in presence of non-cooperative behavior of sensor node in multi-agent WSNs. The tools which are usually employed to select suitable strategies for sensor node in WSNs are non-cooperative game algorithm (Lasaulce and Tembine, 2011) and reinforcement learning (Sutton and Barto, 1998).

1.1.2.1 Non-cooperative game theory

A well-known technique to encourage cooperation among selfish nodes is non-cooperative game theory. Non-cooperative game theory is a branch of game theory which involves interactive decision situations in which

multiple decision makers, each one with its own objectives, jointly determine the outcome. Game theory can be used to analyze the agent interaction and determine a set of strategies among rational agents, where each agent uses available information to decide its behavior. The major advancement that has driven much of the development of game theory is the concept of Nash equilibrium (NE) which is used to determine a suitable and fair strategy for all agents. NE is a set of strategies for each of the agents such that each agent's strategy is the best-response to the other agents' strategies. Many researches focus on the problem of stimulating cooperation. Ref. (Wu and Shu, 2005) applied game theory to routing problem in multi-domain WSNs. They assumed multiple sensor networks under the control of different authorities and used incentive mechanisms to motivate cooperation between sensor nodes. Their approach can be applied in routing and aggregation problems for optimizing the power usage and lifetime of the network. On the other hand, Felegyhazi et al., (2005) applied the Non-cooperative game algorithm to describe a situation that cooperation can exist in multi-domain WSNs without incentive mechanisms. They formulate a packet forwarding game into a non-cooperative resource allocation problem. The authors show that the Non-cooperative game algorithm is a suitable framework which can determine equilibrium strategy for their problem. However, one drawback of these approaches is that obtaining a strategy needs significant amount of computational time to compute the utility for all possible actions of sensor nodes. Similarly, Yang and Brown, (2007) considered co-existing WSNs with two source nodes along with two corresponding destination nodes. A non-cooperative game algorithm is used to analyze the effect of selfishness of sensor nodes on energy efficiency. In their game, each source node acts as agent in a relaying game to send packets to its destination. Each source node

decides to ask or not ask the other source to help relay packets. Their payoff is the amount of energy saved. The results showed that natural cooperation without external incentive mechanisms can occur and can achieve an energy efficiency path selection policy in both fading and non-fading channel. However, their experiment investigated a small network with two sensors and two separate sinks. Moreover, both (Felegyhazi et al., 2005) and (Yang and Brown, 2007) are operated in a centralized manner which are not scalable.

1.1.2.2 Reinforcement learning

In this thesis, we introduce the application of multi-agent reinforcement learning (MARL), another technique to address the issue of resource allocation problem in WSNs. MARL is suitable for distributed routing problems. In the context of reinforcement learning (RL) framework, an agent systematically learns correct behaviors online through trial-and-error interaction with other agent in a dynamic environment in order to achieve a particular goal. There are several recent researches which employ RL to solve routing problems in WSNs (Kulkarni et al., 2011 and Al-Rawi et al., 2015). Each sensor node is assumed to be an agent. Therefore, WSNs with multiple independent decision-making agents can be considered as a in multi-agent reinforcement learning (MARL) system. A standard RL method called, Q-learning has been proposed to determine best routing strategies when critical network conditions are allowed to vary dynamically. In (Yang et al., 2013), a MARL-routing approach was proposed to handle sink mobility and enable direct interactions between WSN and vehicles. Reward functions including time delay, network lifetime and reliability was designed for MARL routing. Simulation results show that their proposed approach achieved better time delay, energy

distribution and delivery rate than often compared routing approaches. Refs. (Hu et al., 2010, Xu et al., 2015 and Debowski et al., 2016) presented a load-balancing multi-path routing approach. A MARL technique was employed to learn the best path to forward packet which considered the number of hops, residual energy and energy consumption of sensor nodes. Results showed that their approach can balance the workload among sensor nodes and prolong the network lifetime. However, these solutions were directly applied to single-domain WSNs.

There are only a few researches which focus on MARL technique in multi-domain WSNs. Ref. (Rovcanin et al., 2014) applied Q-learning to solve routing problem for cognitive networks such networks were co-located heterogeneous WSNs which were fully cooperative and operating in a centralized manner. MARL in a centralized manner was also proposed in (Singsanga et al., 2010), by extending Q-routing to cater a non-cooperative multi-agent in a packet forwarding problem. The authors applied an existing algorithm called Nash Q-learning (NashQ) (Hu and Wellman, 2003) to attain the best mutual policy for all agents in a packet forwarding game. Each agent attempts to learn its Nash equilibrium (NE) online. Their results suggest that NashQ can adaptively learn and determine suitable packet forwarding policy in varying network conditions. However, to the best of our knowledge none of the existing MARL researches take into consideration of fair routing selection in multi-domain WSNs under distributed manner. Since a centralized packet forwarding rely on single computational node to receive and process all sensor data, such operation creates a large amount of overhead rendering it impractical for actual WSN applications (Li et al., 2011). Hence, there is a need for

decentralized or distributed packet forwarding algorithms that allow sensors to estimate their information locally to reduce the amount of overhead used.

This thesis therefore studies the cooperative fair routing problem between multi-domain WSNs which are controlled by different authorities in a distributed manner. The problem of how non-cooperative nodes belonging to different network domains can locally decide to establish cooperative sharing path with other networks without any external incentive mechanisms are taken into consideration. This thesis also studies parameters that effect cooperation between different network authorities and fairness of benefits that the networks can achieve. For this purpose, this thesis focuses on applying MARL and non-cooperative game theory to determine a fair packet forwarding strategy for all network authorities. *The underlying aim of this thesis is to propose a routing algorithm to cater a non-cooperative multi-agent and to achieve the best mutual policy and improve the network performance in distributed multi-domain WSNs.* In order to achieve the aim, this thesis firstly proposes a suitable payoff matrix for packet forwarding game. The payoff matrix is then applied to the proposed *Non-cooperative game algorithm based on Lemke Howson method* (NCG-LH) algorithm to conceptually show that non-cooperative game theory can determine fair packet forwarding strategy and improve the network performance in distributed multi-domain WSNs under common sink (**Chapter 3**) and separate sink scenario (**Chapter4**). This thesis then extends the non-cooperative game theory by adding *learning* mechanism based on *game theoretic reinforcement learning* (GTRL). In particular, the thesis proposes two routing algorithms (**Chapter 5**). The first algorithm is the *discrete state Nash Q-learning* (*D-NashQ*) which is an extension of discrete state NashQ in centralized routing in

(Singsanga et al., 2010) to cater a *distributed* multi-domain WSNs by using the derived payoff matrix as a reward function. The other algorithm is the *continuous state Nash Q-learning* (*C-NashQ*), that considers the state space in the framework as continuous state, which is a suitable representation of the continuous state of the remaining battery energy in the sensor nodes. This thesis also evaluates the proposed algorithms by comparing them to existing algorithms and discusses the network performance. The results show that the proposed algorithms can provide efficient and fair packet forwarding policy that increase the network lifetime and reliability of packet delivery ratio.

To conclude, the main contributions of this thesis are six-fold:

1) Identification of parameters that effect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve.

2) Design of payoff matrix for non-cooperative packet forwarding game in distributed multi-domain WSNs

3) A non-cooperative game algorithm (NCG-LH) is proposed to distributed packet forwarding scheme in non-cooperative multi-domain WSNs under common sink and separate sink scenarios.

4) Proposal of two distributed routing algorithms (D-NashQ and C-NashQ) and their application to the packet forwarding problems in multi-domain WSNs under separate sink scenario.

5) Derivation of feature function that suitable for continuous state Nash Q-learning.

6) Fairness comparison in cooperative routing between routing algorithm based on load balancing technique, non-cooperative game theory technique and game theoretic reinforcement learning technique.

1.2 Research objectives

1.2.1 To identify the parameters that effect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve.

1.2.2 To apply non-cooperative game theory to allocate packet forwarding problem in distributed multi-domain WSNs based on common sink and separate sink scenarios.

1.2.3 To obtain a routing scheme which can achieve the best mutual packet forwarding strategy in non-cooperative multi-domain WSNs in a distributed manner using game theoretic reinforcement learning algorithm.

1.3 Assumptions

1.3.1 Cooperative packet forwarding is beneficial when the network is sparse or when the environment is hostile.

1.3.2 Game theoretic multi-agent reinforcement learning provides more efficiently network performance than the Non-cooperative game approach.

1.3.3 Sensor nodes in multi-domain WSNs can communicate with each other using the same underlying protocol.

1.4 Scope and limitation

1.4.1 Multi-domain wireless sensor network consists of multiple co-located WSNs.

1.4.2 Decision methods for choosing the optimal packet forwarding strategy in multi-domain WSNs will be studied.

1.4.3 Non-cooperative game theory and game theoretic reinforcement learning (GTRL) methods will be studied and compared to achieve a suitable packet forwarding strategy in multi-domain wireless sensor networks.

1.4.4 Simulations will be carried out by Visual C++. Six methods will be compared, namely, 1) AODV non-cooperative routing, 2) AODV cooperative routing, 3) Pool-based routing algorithm (Kinoshita et al., 2016) 4) the proposed method on Non-Cooperative Game based on Lemke Howson (NCG-LH) method, and the proposed method on game theoretic reinforcement learning algorithms namely, 5) Discrete state Nash Q-learning (D-NashQ); and 6) Continuous state Nash Q-learning (C-NashQ). The experimental results will be analyzed to find the suitable and fair packet forwarding strategy.

1.5 Expected usefulness

1.5.1 A game theoretic multi-agent reinforcement learning algorithm can be applied to find the best mutual policy for packet forwarding in non-cooperative multi-domain WSNs.

1.5.2 An optimal and fair packet forwarding strategy for non-cooperative multi-domain wireless sensor networks.

1.6 Synopsis of thesis

The remainder of this thesis is organized as follows. **Chapter 2** presents the theoretical background which is the foundation for the contributions of this thesis. Firstly, the concept of non-cooperative game theory formulation and NCG-LH

algorithm are introduced. Secondly, the concept of the Markov decision process formulation is reviewed. Next, game theoretic reinforcement learning technique used for solving the packet forwarding problem called D-NashQ and C-NashQ algorithms are introduced.

Chapter 3 presents a suitable payoff metric for packet forwarding game and conceptually show that NCG-LH algorithm can be applied to allocate packet forwarding problem in distributed multi-domain WSNs based common sink scenario.

In **Chapter 4**, the packet forwarding game is formulated and solved by the NCG-LH algorithm for resource allocation problem between multi-domain WSNs in separate sink scenario.

Chapter 5 proposes the game theoretic reinforcement learning techniques called D-NashQ and C-NashQ algorithms in multi-domain WSNs. The packet forwarding game was formulated and solved by D-NashQ and C-NashQ algorithms.

Finally, **Chapter 6** summarizes all the original findings and contributions in this thesis and points out possible future research directions.

CHAPTER II

BACKGROUND THEORY

2.1 Introduction

This thesis studies the cooperative fair routing problem in multi-domain wireless sensor networks (WSNs). An important usage for multi-domain WSNs is resource sharing between different authorities which can prolong their lifetime. However, cooperative behavior between sensor nodes belonging to different authorities may not always be readily available because sensor nodes may act selfishly to conserve their energy. Furthermore, there is no guarantee that node cooperation will be beneficial to all WSNs. Therefore, it is necessary to find an algorithm for each authority to decide whether to cooperate with each other or not in a non-cooperative multi-domain WSN.

This thesis applies non-cooperative game theory and reinforcement learning (RL) to address the issue of non-cooperative resource allocation problem in multi-domain WSNs. Non-cooperative game theory (Shoham and Brown, 2009) analyzes the interaction and determine a set of strategies among rational selfish agents, where each agent uses available information to decide its behavior for a given outcome. On the other hands reinforcement learning (RL) (Sutton and Barto, 1998) is a machine learning scheme to provide a framework in which an agent learn the optimal policy based on the agents' past experiences without full information about the model of the

environment. Non-cooperative game theory and RL thus are employed to encourage cooperative fair routing between sensor nodes in multi-domain WSNs.

Therefore, this chapter serves as an introductory to important concepts of game theory and then the fundamental theory of reinforcement learning which are the basis of the contribution of this thesis.

2.2 The Agent definition

This thesis focus on the problem of packet forwarding cooperation in multi-domain WSNs. In particular, a packet forwarding game is formulated into a non-cooperative resource allocation problem. The term 'agent' in this thesis represents a decision maker which decides an optimal route to forward the data packet. We assume that the agent is rational, if given what the agent knows so far, the agent will always choose a strategy which optimizes some performance measure.

In this thesis, the source node takes a role as an agent. The source node is randomly selected from the set of sensor nodes in the WSN to send packets to the base station (or sink node). The source node needs to decide which route obtains the best benefit for its network domain. This thesis models a packet forwarding game as a two-agent game. The example of the game is shown in Figure 2.1. From the figure, source node n_1^1, which is randomly selected from sensor nodes in network domain1, is modeled as an agent in the game. The agent n_1^1 must decide whether to use the non-cooperative route which uses nodes in its own domain or the cooperative route that consists of nodes from the other domain. To make a decision, the agent n_1^1 assumes that neighbor node n_2^1, which is a sensor node in a different network domain

(domain2) belonging to cooperative route, as the other agent in game. The other agent's behaviors is expected to act rationlly.

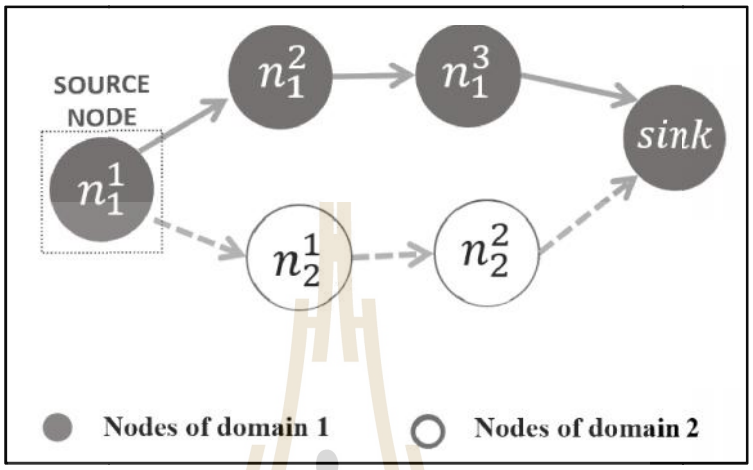


Figure 2.1 Example of packet forwarding game

The optimal packet forwarding route is chosen by the source node depending on strategy decision obtained from the proposed algorithms described in Chapter 3, 4 and 5.

2.3 Non-cooperative game

In recent years, game theory has gained much gaining attention in wireless network researches because as it is powerful to analyze rational agent (or player) behavior (Lasaulce and Tembine, 2011). Game theory has been successfully applied in a wide range of problems spaces such as data routing (Fan et. al., 2016), power control (Al-Zahrani et. al., 2016), wireless security system (Wang et. al., 2014) and intrusion detection (Moosavi and Bui, 2014). Non-cooperative game theory is a branch of game theory which involves interactive decision situations in which

multiple decision makers, each one with its own objectives, jointly determine the outcome of the decision.

In this thesis, Non-cooperative game theory has been applied in cooperative routing problems, which is usually referred to as packet forwarding game, for multi-domain WSNs. The idea behind the employment of non-cooperative game theory in routing area is that the agents, e.g., sensor nodes in WSNs, which have a rational selfish behavior, attempt to benefit themselves first when they are making packet forwarding decisions. Thus, these selfish sensor node may prefers to drop a packet from other different network domain rather than help to forward for conserving limited energy resources since each data packet transmission has a cost for each sensor node that participates in the route. The cooperative routing between multi-domains can be broken if all nodes in different domain adopt this strategy. Under such scenario, each agent needs to consider other agents' benefits while optimizing its own benefits in making decisions in order to avoid failure in cooperation. Non-cooperative game theory is capable of providing a set of mathematical tool to analyze such complex interactions among rational selfish agents.

2.3.1 Game strategic form

Strategic form (or normal form) is a basic component in game theory, which is defined by the tuple, (I, A, u) where

- I denotes the set of agents, $i \in I$, $i = 1, ..., I$
- $A = A_1 \times ... \times A_I$, where A_i is the set of actions available to agent i, and tuple $(a_1, ..., a_I) \in A$ is called an *action profile*, which describes the action each agent has chosen.

- $u = (u_1, ..., u_I)$, where u_i is a real-valued payoff function for agent i.

This thesis refers agent i's opponents as "-i". Note that we consider strategic games with complete information, meaning that each agent has knowledge about all the other agents' payoff functions.

Appropriate *strategies* for the game can be determined by the application of *solution concepts*, which determine. In other words, solution concepts can determine what strategies for agents are suitable to adopt in the game. The most widely used solution concept is Nash Equilibrium (NE). The next section, we will describe concept of NE and method to find NE.

2.3.2 Nash equilibrium concept

In game theory, the Nash equilibrium (NE) is a solution concept of a non-cooperative game which is used to determine a suitable and fair strategy for all agents. NE is a set of strategies for each of agent such that each agent can correctly expect about of the other agent's behaviors, and acts rationally to this expectation. Acting rationally signifies that the agent's strategy is the best response to the other agents' strategies. For any game, NE is at least one solution exists in pure or mixed strategies (Sutton and Barto, 1998). Given a set of strategies, if the agents choose to take their action with probability 1, this implies that the agent is playing in a pure strategy. On the other hand, a mixed strategy is a probability distribution over pure strategies. The agents need to select their action according to some probability.

In pure strategy NE, an agent selects an action which achieves the best response to the other agent's choice. In other words, a pure strategy NE is a point of joint strategy in the stage game which every agent receives its highest payoff at this point, and a change in strategies by any one of them would result in lower gains for

that agent than the current strategy. Mathematically, the *strategy profile* $a_1^*,...,a_I^*$ is a NE if for all agent *i*, a_i^* is the best response to the other agents' choices a_{-i}^*

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*) \tag{2.1}$$

where $u_i(a_i, a_{-i})$ is payoff for agent *i* received after choose joint action (a_i, a_{-i}) and $a_i \in A_i$

In general, the existence of a pure NE for the game cannot be guaranteed. However, a mixed strategy NE always exists in finite games. Therefore, it is necessary to extend the concept of NE to include mixed strategy NE in order to analyze for solutions.

2.3.3 Generating Nash equilibrium using Lemke-Howson method

In this section, we will consider mixed strategy NE, which exists for every finite game. A mixed strategy is a strategy in which an agent performs its available pure strategies with certain probabilities. A mixed strategy NE profile $\pm_1^*,...,\pm_I^*$ is a NE if for every agent *i*, $\pm_i^*$ is the best response to the other agents' choices $\pm_{-i}^*$,

$$u_i(\pm_i^*, \pm_{-i}^*) \geq u_i(\pm_i, \pm_{-i}^*) \tag{2.2}$$

for each $\pm_i \in \Sigma_i$,when Σ_i is the probability distribution over agent *i*'s pure strategies.

In this thesis, the Lemke-Howson (LH) method (Sutton and Barto, 1998) is employed to calculate the probability to achieve the NE in a Non-cooperative game. The LH method is the best known method to solve for mixed-strategy NE between two agents. The advantage of LH method is that it is guaranteed to find at

least one NE point. More details about using LH method to finds NE is shown in Appendix A.

2.4 Multi-agent online learning approach

Reinforcement learning (RL) (Sutton and Barto, 1998) is a machine learning scheme in which an agent learns the optimal policy from the agents' past experiences without prior information about the model of the environment. Convergence of RL relies on the assumption that the dynamics of environment satisfies a Markov Decision Process (MDP). Therefore, this section start with a theoretical background on MDP theory follow by a description of reinforcement learning and Nash Q-learning.

2.4.1 Markov decision process theory

A Markov decision process (MDP) is the foundation for single-agent reinforcement learning. MDP provides a framework for modelling that consists of a decision-maker interacting synchronously with a signal from the environment called the environment's *state*. If the decision-maker sees the environment's true state, it is referred as a *completely observable Markov decision process*. The foundation of MDP is presented as follows.

2.4.1.1 Markov property

The Markov property states that anything that has happened so far can be summarized by the current state. Thus, the probability of being in the next state at time t+1 based on the past history of state changes can be defined simply as the conditional probability based on the current state at time t,

$$P(S^{t+1} = s^{t+1} \mid S^t = s^t, ..., S^0 = s^0) = P(S^{t+1} = s^{t+1} \mid S^t = s^t). \quad (2.3)$$

This equation is referred to as the Markov property. A state refers to information on the environment that may be useful in making a decision. If the state has the Markov property, then the environment's state at time t+1 depends only on the state representation at time t.

2.4.1.2 Markov Decision Process

A MDP is a discrete-time random decision process defined by a set of states, actions and the one-step dynamics of environment. Given any state s and action a, the probability of occurrence of each possible next state $s^{t+1} = s'$ is

$$P(s' \mid s, a) = P(S^{t+1} = s' \mid S^t = s, a^t = a). \quad (2.4)$$

This equation is called the state transition probability. Similarly, given any current state and action, s and a, together with any next state, s', the expected value of the incurred reward is

$$R(s, a, s') = E[r^{t+1} \mid S^t = s, a^t = a, S^{t+1} = s'], \quad (2.5)$$

where $E[\cdot]$ is the expectation operator and r^{t+1} is the reward received at time t +1. Equation (2.4) and (2.5) completely specify the most important aspects of the dynamics of the MDP. A MDP model is shown in Fig. 2.2.

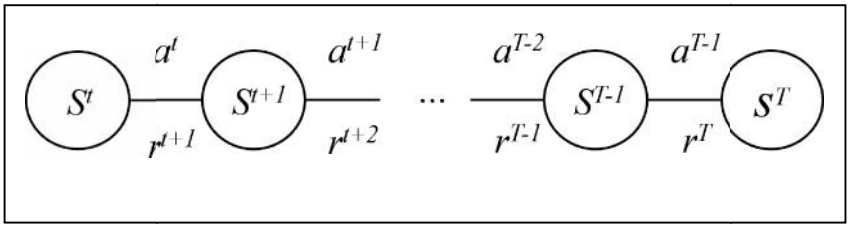


Figure 2.2 A MDP model.

A tuple (*S, A, P, R*) is used to describe the MDP characteristics, where *S* is the discrete set of environment states, *A* is the discrete set of possible actions. In each time step *t*, the agent will observe the current state $s^t = s \in S$ and select an action $a^t = a \in A$. After taking action, the environment makes a transition into a new state $s^{t+1} = s' \in S$ according to the state transition probability $P(s' \mid s,a) \in P$ and then receives a feedback $r^t \in R$ which is a function of the reward expected from the environment as a result of taking action $a \in A$. Let π be defined as a mapping of the state space to the action space, $\pi : S \to P[A]$, where *P[A]* is the distribution over the action space. The objective of solving a MDP is to find a policy π that maximizes (or minimizes) some desired objective function. Such objective function is defined as follows. Let $Q_t^{\pi}(s,a)$ be defined as the action-value function of a given policy π which associates state-action pair (s,a) with an expected reward for performing action *a* in state *s* at time step *t* and following π thereafter;

$$\begin{aligned} Q^{\pi}(s,a) &= E^{\pi}[R^t \mid s^t = s, a^t = a] \\ &= E^{\pi}[\sum_{k=0}^{\infty} r^{t+k+1} \mid s^t = s, a^t = a], \end{aligned} \tag{2.6}$$

where $R^t = r^{t+1} + \quad r^{t+2} + \quad ^2 r^{t+3} + \ldots = \sum_{k=0}^{\infty} \quad ^k r^{t+k+1}$ is the expected discounted return of the agent, is the discount factor and $E^{\pi}[\cdot]$ is the expectation operator under policy π.

The objective of MDP is to find a policy to select actions at a given state such that the long term average reward is maximized. To achieve this, particularly in scenarios where the dynamics of the environment is difficult to model

(such as in WSNs), a technique called reinforcement learning can be used to solve MDPs.

2.4.2 Reinforcement learning

Reinforcement learning (RL) is a computational approach which identifies how a system in a dynamic environment can learn to choose optimal actions to achieve a particular goal. The learner is not taught which action to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trial-and-error interactions with its environment (Sutton and Barto, 1998).

In RL model, the learner or decision maker is called the agent. Everything outside the agent is called environment. It uses a formal framework defining the interaction between a learning agent and its environment in terms of states (s^t), actions (a^t) and rewards (r^t). The agent selects actions and the environment responds to those actions. Furthermore, the environment also feed backs to the agent rewards, as a consequence of the action selection at a given state, which the agent tries to maximize over time. More specifically, the agent and environment interact with each other in a sequence of discrete time steps. At each time step (t), the agent receives some representation of the environment's state (s^t) and selects an action (a^t). One time step later, the agent receives a numerical reward (r^{t+1}) and finds itself in a new state (s^{t+1}). Figure 2.3 shows the agent-environment interaction in reinforcement learning.

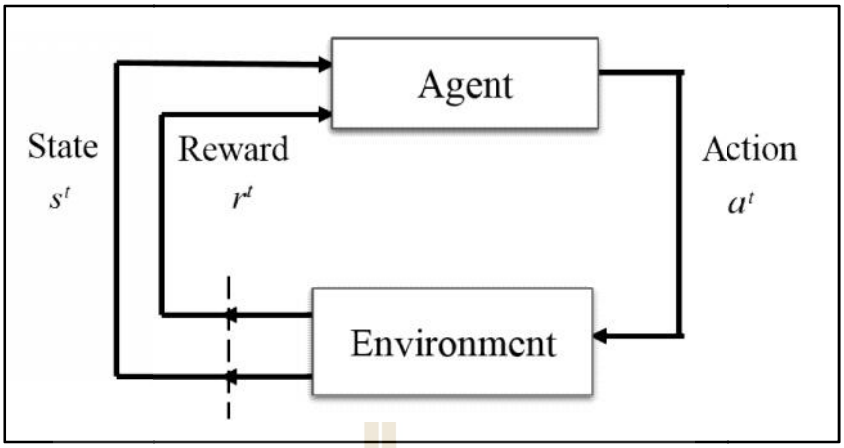


Figure 2.3 Diagram of agent-environment interaction in reinforcement learning.

2.4.2.1 The value function

Reinforcement learning algorithms are based on estimating value functions. A value function is the expected sum of rewards received from starting in state *s*. The value functions evaluate the performance of the decision which the learner has taken at a given state. Since the rewards received in the future by the learner depend on the actions which are taken, value functions are defined with respect to each particular policy. Therefore, we can define the value function of a state under a policy® , $V^{®}(s)$, as

$$V^{®}(s) = E^{®}[R^t \mid s^t = s]$$

$$= E^{®}[\sum_{k=0}^{\infty} {}^{k} r^{t+k+1} \mid s^t = s], \tag{2.7}$$

where $E^{®}[\cdot]$ is the expectation operator under policy ®. We call function $V^{®}$ the state-value function.

In RL, the agent attempts to improve its decision-making policy® over time in order to learn an optimal policy $®^*(s)$ for each state *s*, which is maximize the total expected discounted reward over the long run. The optimal state-value function, denoted as $V^*(s)$, would therefore be the state value function which is maximum over all possible policies at state *s*.

$$
\begin{aligned}
V^*(s) &= V^{®^*}(s) \\
&= \max_{®} V^{®}(s) \qquad (2.8) \\
&= \max_{®} E^{®}[\sum_{k=0}^{\infty} {}^{k} r^{t+k+1} \mid s^t = s] \\
&= \max_{®} E^{®}[\mathrm{r}^{t+1} + \sum_{k=0}^{\infty} {}^{k} r^{t+k+2} \mid s^t = s] \\
&= \max_{®} E^{®}[r^{t+1} + V^*(s^{t+1}) \mid s^t = s] \\
&= \max_{a} \sum_{s'} P(s' \mid s,a)[R(s,a,s') + V^*(s')] \\
&= \max_{a} \left[R(s,a) + \sum_{s'} P(s' \mid s,a) V^*(s') \right], \qquad (2.9)
\end{aligned}
$$

where $P(s' \mid s,a)$ is the probability of transiting to next state s' after taking action *a* at state *s*. The quantity $R(s,a)$ is the expected next reward given the current state and action, that is $R(s,a) = E[r^{t+1} \mid s^t = s, a^t = a]$, and is related to $R(s,a,s')$ by $R(s,a) = \sum_{s'} P(s' \mid s,a) R(s,a,s')$. Equation (2.9) is called the *Bellman optimality equation* for V^*. This equation is also known as *iterative policy evaluation* (Puterman, 1994).

However, in many situations the state transition probability and reward model in (2.9) is unknown. Therefore, such models can be learnt directly by an agent interacting directly with the environment. Such approach is called *model-free reinforcement learning*. One popular model free reinforcement learning technique used in this thesis is presented next.

2.4.2.2 Q-learning

Q-learning (Sutton and Barto, 1998) defines a learning method within a MDP that is employed in single-agent RL systems. Q-learning is an algorithm that does not need a model about the state transition probability and can directly approximate the optimal *action-value function* (*Q-value*) through online learning. We can define the right-hand side of Eq. (2.9) by

$$\begin{aligned} Q^*(s,a) &= Q^{\circledR *}(s,a) \\ &= R(s,a) + \sum_{s'} P(s' \mid s,a) V^*(s') \end{aligned} \tag{2.10}$$

where $Q^*(s,a)$ is the total discounted reward of taking action a at state s. Then, we obtain

$$V^*(s) = \max_a Q^*(s',a'). \tag{2.11}$$

It can be seen that the optimal value function $V^*(s)$ is substituted into (2.10), we can write the $Q^*(s,a)$ as a function of $Q^*(s',a')$ as follows.

$$Q^*(s,a) = R(s,a) + \sum_{s'} P(s' \mid s,a) \max_{a'} Q^*(s',a') \tag{2.12}$$

Eq. (2.12) is called *Bellman optimality equation* for Q^*.

In Q-learning, the agent tries to find the optimal $Q^*(s,a)$ by iteratively updating the estimate $Q(s,a)$. The basic idea in Q-learning is to estimate

Q-value for actions based on feedback (reward) and agent's Q-value function using the observed information $\langle s, a, r^t, s' \rangle$. The update rule is based on *temporal-difference* (TD) learning, which using Q-values and the Q-learning estimated from the next state in order to update $Q^t(s,a)$ to $Q^{t+1}(s,a)$. Q-learning provides a simple updating process, in which the agent starts with an arbitrary initial Q-value $Q^t(s,a)$ for all $s \in S, a \in A$. After executing action *a* at state *s*, the agent receives an immediate reward *r* and then transits to a new state s' and updates the new Q-value at time step *t+1* as follows :

$$Q^{t+1}(s,a) = Q^t(s,a) + \sim^t [r^t + \max_{a'} Q^t(s',a') - Q^t(s,a)]$$

$$= (1 - \sim_t) Q^t(s,a) + \sim^t [r^t + \max_{a'} Q^t(s',a')], \qquad (2.13)$$

where $\sim^t \in [0,1)$ is the learning rate and $\in [0,1)$ is the discount factor. The process is repeated so that the agent can learn its own optimal policy. Note that the Q-value in equation (2.13) can converge to $Q^*(s,a)$ under the assumption that all states and actions have been visited infinitely often. The optimal policy is given by

$$®^*(s) = \max_a Q^*(s,a). \qquad (2.14)$$

It can be seen that Q-learning provides a simple procedure to learn optimal policy in single-agent RL systems.

2.5 Multi-agent in non-cooperative game

Multi-agent systems differ from single-agent systems in that there are many different agents that are supposed to learn a task and that all of the agents' actions

affect the environment. Thus, each agent needs to maintain observation of its environment and as well as the other agents in order to learn the optimal policy. Therefore, the Q-learning algorithm for single agent is extended to consider other agent' actions as well.

The author in (Hu and Wellman, 2003) proposed the Nash Q-learning (NashQ) algorithm, by extending Q-learning to a non-cooperative situation where each agent can rationally decide its action whether it will cooperate with other agents or not by considering both its own and other agents' information as well.

2.5.1 The action-value function

Instead of finding an optimal action to maximize one single agent's reward as the single-agent Q-learning, NashQ seeks joint actions that yield the best possible reward for all agents. For a *two-agent system*, the action-value function for agent *i* becomes $Q_i(s_i, a_1, a_2)$, where *i*=1,2.

The objective of the agents in the NashQ algorithm is to learn their best mutual response policy, which is defined by the Q-values received from Nash equilibrium (NE). NE is not only used to decide the agent's own action policy, but also predict the other agent's action, given by $®_1(s_1'), ®_2(s_2')$ where $®_i(s')$ is agent *i*'s distribution over the set of actions at state s'. NashQ then calculates a NE for the stage game $(Q_1^t(s'), Q_2^t(s'))$ and updates its Q-values according to

$$Q_i^{t+1}(s_i, a_1, a_2) = (1 - \sim^t) Q_i^t(s_i, a_1, a_2) + \sim^t [r_i^t + \quad NashQ_i^t(s_i', a_1', a_2')], \tag{2.14}$$

where (a_1, a_2) is a joint action, r_i^t is an immediate reward for agent *i* in the state *s* under this joint action and $NashQ_i^t(s_i', a_1', a_2')$ is the Q-value of agent *i* in the next state s_i' for selecting joint action received from Nash equilibrium, which is defined by

$$NashQ_i^t(s_i', a_1', a_2') = ®_1(s_1') \cdot Q_t^i(s_i', a_1', a_2') \cdot ®_2(s_2') \tag{2.15}$$

In order to calculate the Nash equilibrium, agent *i* must observe the other agent's immediate reward and previous actions and updates its conjectures on the other agent's Q-value, by maintaining its own update on the other agent's Q-value:

$$\begin{aligned} Q_j^{t+1}(s_j, a_1, a_2) = (1 - \sim^t) Q_j^t(s_j, a_1, a_2) \\ + \sim^t [r_j^t + \quad NashQ_j^t(s_j', a_1', a_2')], \quad j \neq i. \end{aligned} \tag{2.16}$$

NE can be found in a pure-strategy equilibrium, where an agent is able to achieve the best response to the other agent's choice. However, not all games have pure-strategy equilibrium (Daskalakis et. al., 2009). Under this circumstance, the agents need to select their strategies randomly according to some probability calculated from the Lemke-Howson method (Shoham and Brown, 2009) to achieve the NE. Such equilibrium is called mixed-strategy equilibrium (see Appendix A for more details).

Initialize:

Let $t = 0$, get the initial state s_0.

Let the learning agent be indexed by i.

For all $s \in S$ and $a_i \in A_i, i = 1,2$, let $Q_i^t(s_i, a^1, a^2) = 0$.

Loop

Choose action a_i^t.

Observe $r_1^t, r_2^t; a_1^t, a_2^t$, and $s^{t+1} = s'$

Update Q_i^{t+1} for i=1,2

$Q_i^{t+1}(s_i, a_1, a_2) = (1 - \sim^t) Q_i^t(s_i, a_1, a_2)) + \sim^t [r_i^t + \quad NashQ_i^t(s_i', a_1', a_2')]$,

where $NashQ_i^t(s_i', a_1', a_2')$ is defined in (2.15)

Let $t := t + 1$.

Figure 2.3 The Nash Q-learning algorithm (Hu and Wellman, 2003).

2.5.2 Convergence

NashQ requires two conditions in a stage game during learning to converge (Hu and Wellman, 2003).

1) The stage games encountered during learning have a global optimal point, which is defined as a point of joint strategy in the stage game which every agent receives its highest payoff at this point, or

2) They all have a saddle point which is defined as a point of joint strategy in the stage game which is a NE point, and each agent would receive a higher payoff when at least one of the other agents deviates.

However, both the global and saddle points may not always be satisfied for these conditions because of both points may not exist in every stage game. Another limitation is that in selecting NE under a mixed strategy, NashQ algorithm resorts to a mixed strategy selection where the Nash equilibrium is

probabilistically selected according to the Lemke Howson method (Shoham and Brown, 2009). Their algorithm showed that convergence can still be established with such relaxed convergence conditions.

2.6 Summary

In this chapter, an overview of the non-cooperative game theory and the multi-agent Q-learning algorithm called NashQ are given. Both algorithms are used to determine the packet forwarding strategies in non-cooperative multi-domain WSNs in this thesis. By considering joint actions, the agents can rationally determine the best mutual policy and receive fair benefit for all agents in multi-domain WSNs.

In the next chapter, a packet forwarding formulation in non-cooperative multi-domain WSNs is presented. Non-cooperative game theory based Lemke Howson method is used to study the conditions which equilibriums can exist and its performance is evaluated under common sink scenario.

CHAPTER III

PACKET FORWARDING IN COMMON SINK MULTI-DOMAIN WIRELESS SENSOR NETWORKS USING NON-COOPERATIVE GAME

3.1 Introduction

Routing has been a challenging issue addressed in wireless sensor networks (WSNs) mainly due to the scarcity of energy and on-board resources. In recent years, applications of large scale WSNs are becoming a reality. Examples include smart grids (Zaballos et. al., 2011; Fadel et. al., 2015), the Internet of Things (Mattern et. al., 2011; Mulligan 2010) and Machine-to-Machine (M2M) communications networks (Fan et. al., 2011; Niyato et. al., 2011). It is possible that multiple sensor networks can coexist independently within a region of interest without conflicting each other. These networks may even be physically overlapping and their sensor nodes may be interleaved. Such networks are referred to as multi-domain wireless sensor networks (WSNs). These networks could potentially gain certain benefits, such as alternative routing paths and reduced energy consumption, if their sensor nodes share resources which can prolong their lifetime. Many existing works consider resource allocation problems in multi-domain WSNs (Shamani et al., 2013; Jelicic et al., 2014; Singhanat et al., 2015; Kinoshita et al. 2016). All of these works showed that resource sharing and fully cooperation between multiple different networks, result in reduced energy

consumption and increased network lifetime. However, because of possible selfish behaviors among sensor nodes to conserve their energy, cooperation between sensor nodes belonging to different network authorities may not always be readily available. Furthermore, it is also possible that, under certain situations, node cooperation will not be beneficial to any network in the multi-domain WSN. Vaz et al., (2008) and Ze et al., (2012) showed that cooperation between two different networks that are deployed in the same region may not always be beneficial to both networks. This is because whether or not each agent will cooperate depends on the configuration of each network, network connectivity and how hostile the environment is. Previous works have proposed a centralized packet forwarding scheme in Non-cooperative multi-domain WSNs. However, the centralised operation is not scalable. In this thesis, our focus is thus on determining a *distributed* resource allocation scheme for Non-cooperative sensors in multi-domain WSNs which allow each individual sensor to decide its packet forwarding strategy in a distributed manner allowing a more scalable implementation.

In this chapter, we introduce the application of game theory to address the issue of Non-cooperative distributed resource allocation problem in multi-domain WSNs. In particular, game theory can be used to analyse the interaction and determine a set of strategies among rational agents, where each agent uses available information to decide its behaviour. The major advancement that has driven much of the development of game theory is the concept of Nash equilibrium (NE) which is used to determine a suitable and fair strategy for all agents (AlSkaif et. al., 2015). NE is a set of strategies for each of the agents such that each agent's strategy is the best-response to the other agents' strategies. In a game where there is only a single unique

NE, the game is said to have a pure strategy form. However, there are games where no pure NE exists. In such games, there may not be any pure strategy that provides the maximum payoff for all agents (i.e., an agent always attains higher payoff than other agents). Therefore, in such games, each agent can choose its pure strategy with a certain probability, which results in a (probabilistic) *mixed strategy equilibrium*.

Many researches focus on the problem of stimulating cooperation between WSNs. Ref. Wu et al. (2005) and Miller et al. (2005) applied game theory to packet forwarding in multi-domain WSNs problems by using incentive mechanisms to motivate cooperation between sensor nodes. Incentive mechanism such as using trust values are used to encourage cooperation packet forwarding among nodes. On the other hand, (Felegyhazi et al., 2005) and (Yang et al., 2007) applied the Non-cooperative game algorithm to determine a situation which cooperation can exist in multi-domain WSNs without any incentive mechanisms. Cooperation may exist only when it achieves mutual benefits for every network. The rational for this is that cooperation is advantageous in certain situations when the payoff exceeds the actual costs of cooperation. Therefore, there is no need to use incentives to cooperate in every situation. Ref. (Felegyhazi et al., 2005) formulated a packet forwarding game as a Non-cooperative resource allocation problem. They showed that the Non-cooperative game algorithm is a suitable framework to determine an equilibrium strategy for their problem. However, this algorithm requires a centralised operation to determine the packet forwarding strategy for each agent (in a centralized operation, an agent refers to the cluster head in each network). Moreover, due to sensor nodes' communication and energy constraints, a centralized payoff estimation, in which a single computational node receives all sensor data, is inefficient and not scalable. The

global information maintained by each agent creates a large amount of overhead. Hence, a decentralized or distributed algorithms that allow sensor nodes to estimate their payoff locally to reduce the amount of overhead used would be more practical. Ref. (Yang et al., 2007) considers the problem of relay selection in a packet forwarding problem in multi-domain WSNs with selfishly behaving nodes. A payoff matrix is implemented to compare the amount of energy a node can save. A NE strategy is then selected based on the payoff matrix. Although their results show that NE can indeed achieve cooperation, their work is based on a small network with a single relay node in each network. In practice, a network consists of several tens, hundreds or even thousands of sensor nodes. Furthermore, the payoff matrix used in (Yang et al., 2007) did not take into consideration the packet receiving rate (PRR) despite the fact that their relays must satisfy SNR constraints.

This chapter therefore studies packet forwarding problem between sensor nodes belonging to multi-domain under Non-cooperative and hostile conditions. For this purpose, we propose a novel *payoff* matrix and propose the *Non-cooperative game algorithm* to determine the best packet forwarding strategy for all network authorities in the system. It is worth noting that this thesis considers a *localised distributed* approach, as opposed to the centralised approach in (Felegyhazi et al., 2005), to reduce the amount of communication overhead and achieve scalability. The proposed payoff matrix in this thesis differs from (Yang et al., 2007) in that it takes in to consideration of successful packet delivery in terms of the packet reliability ratio, in addition to the energy savings.

The underlying objective of this chapter is propose a novel payoff matrix to determine a mutual strategy for the packet forwarding problem in a Non-cooperative multi-domain WSNs which enables cooperation in necessary network environments to achieve packet reliability and to prolong network lifetime as mutual benefits for all domains. Non-cooperative game algorithm is applied to decide a suitable course of action for the agents in the packet forwarding game. This chapter will also study the NE conditions of the packet forwarding strategies in multi-domain WSN and fairness issues in terms of the energy usage in each domain. In situations where there is no pure strategy, the well-known Lemke Howson method is used to determine a mixed strategy for games with two agents (Shoham and Brown, 2009). To evaluate the performance of Non-cooperative game algorithm, we divide the experiment into two parts. In the first part, we formulate our packet forwarding game into uniform random topology framework in order to show that Non-cooperative game theory can be applied to obtain the best mutual policy in small scale WSNs. The second part extends the study to a more realistic scenario by replacing Non-cooperative game theory to tree topology WSNs.

The main contribution of this chapter is three-fold: 1) the distributed packet forwarding scheme in Non-cooperative multi-domain WSNs; 2) identification of parameters that effect cooperation between multi-domain networks and fairness of benefits that the networks can achieve; 3) a novel payoff matrix to be used in packet forwarding in Non-cooperative multi-domain WSNs.

3.2 Non-cooperative game

3.2.1 Game theoretic framework

Non cooperative games are game situations consisting of at least two agents whereby the decision making of an agent involves knowledge of the interactions or strategies from other agents in the game. Each agent is considered rational which would undertake actions to gain its own maximum benefits or payoffs. Each agent independently selects its own action without any prior negotiation which makes it suitable for non-cooperative behavior in multi-domain WSNs. If a sensor node has a packet to send to the sink node, that sensor node becomes a source node. Each source node takes a role of an agent in the game which acts selfishly to conserve their limited energy supply. Each agent makes its own decision for the maximum benefit or payoff for its own network.

3.2.2 Packet forwarding game using Non-cooperative game approach

In Non-cooperative Game, each agent can independently decide to interact with the other agents without any prior agreement or collaborative conditions. Therefore, it is necessary for each agent to predict actions of other agent in order to determine its own action, relative to the others. The Non-cooperative game algorithm (Lasaulce and Tembine, 2011), is a branch of game theory applied exclusively to the situation where the interests of multiple agents conflict. Such situation may arise in a multi-domain WSNs, where sensor nodes may wish to forward packets using nodes from the other domain to conserve their own energy. A basic component in non-cooperative game is defined by the tuple, (I, A, u), where I denotes the set of agents, A denotes the set of actions (i.e. policies) and u denotes a set of payoff functions. The solution in the Non-cooperative game algorithm is based on Nash equilibrium (NE)

which attains the best mutual policy for all agents in the game. The following notations are defined for a game:

1) *Agents* refer to source nodes in each network. The source nodes must make decisions selecting a route to forward a packet to the base station.

2) *Action* refers to the set of possible actions which can be selected by the agents. In this research, there are two actions which agents (source) nodes can select, i.e., a Non-cooperative route or a cooperative route. A Non-cooperative route comprises nodes with in the same domain only whereas a cooperative route consists of nodes from other domains as well. Source nodes make their decisions upon the NE from a matrix of payoff functions which each sensor node maintains.

3) *Payoff function* is the outcome resulting from the agents' interaction according to the selected action. It can be defined in terms of energy savings, energy consumption or packet delivery.

This chapter proposes a payoffs matrix by improving that in (Yang and Brown, 2007) by not only considering energy savings for each strategy, but also taking into consideration the packet receiving rate (PRR) or the packet delivery rate (Ahmed and Fisal, 2008). The payoff matrix is then used in the Non-cooperative game presented in the following section.

3.3 Problem formulation

In this section, Non-cooperative game algorithm is formally introduced in order to find the best mutual policy for packet forwarding in multi-domain WSNs.

3.3.1 Packet forwarding game

In our model, we assume that there are two different networks in the multi-domain WSN. Let N_i be the set of nodes in each network such that $N_i = \{n_i^1, n_i^2, \dots, n_i^v\}$, where *v* is the number of sensor nodes in network *i*. We assume that the system operates as a distributed system whereby a source node in each WSN determines its own behaviour and decides its packet forwarding strategy independently. The role of each sensor node in multi-domain WSNs is to send its data measurements (i.e., packets) to neighbouring nodes through multi-hop communication to a common base station. We assume that two sensor nodes are able to communicate when they are within transmission range. Even if sensor nodes belong to a different network, interactions between the agents are assumed. It is also assumed that in each network, an AODV path discovery scheme is used. When there is a packet to be sent to the common base station, the source node broadcasts a RREQ message to its neighbouring sensor nodes in the same domain, which will in turn broadcast the message to their neighbours until non-cooperatives route to the sink are discovered. In a similar manner, the source node also discovers cooperative routes by broadcasting RREQ messages to sensor nodes belonging to the other domain as well. Therefore, each agent maintains two different routing tables, one for routing within their own network and the other for routing through coordinated paths with the other network. The shortest route in the two tables are selected as the action for the source node (i.e., the Non-cooperative and cooperative route). Each route incurs an energy cost associated to it described as follows (as shown in Figure 3.1). From the figure, let a sensor node from network domain 1, n_1^j , where *j=1,2,..v,* be a source node taking a role as an agent in the game which has a packet to send to its sink1. Suppose n_1^j has to

make a decision whether to use the action that uses the non-cooperative route in its own domain or the action that uses the cooperative route that consists of nodes from the other domain (i.e. forward its packet to n_2^k, where *k=1,2,..v*). The decision to select which action is based on the energy model in section 3.3.2 and the strategy decision model in section 3.3.3

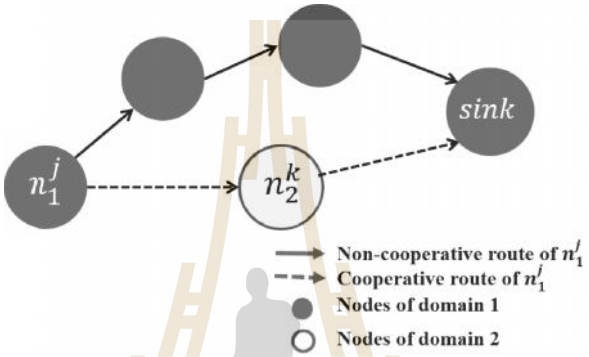


Figure 3.1 System model for common sink

3.3.2 Radio model

The energy consumption required for packet forwarding is computed from the radio model in (Naruephiphat and Usaha, 2008). The radio model for the reception cost of agent n_i^j is given by, $E_{i,RX}^j(b) = E_{elec} \times b$ where E_{elec} is the cost in the radio electronics E_{elec} = 50 nJ/bits and we assume that *b* is the size of the measurement packet transmitted in bits. Therefore, the transmission cost of agent n_i^j is given by, $E_{i,TX}^j(b,d) = E_{elec} \times b + (\mathcal{E}_{amp} \times b \times d^{\pm})$ where σ is the path loss exponent and ε_{amp} is the energy consumed at the output transmitter antenna for transmission range of one meter, $\varepsilon_{amp} = 10pJ/bit/m^2$. We assume that the agent sends its packet to a common base station (called sink) by either its own route or a coordinated route through the

other network depending on action selected by the agent. The strategy decision is presented next.

3.3.3 Action decision

In the Non-cooperative game, each agent can independently decide its own action whether or not to cooperate with the other agent. A set of actions, which include all the possible joint actions available in the game, is defined by $A_i = \{D, F\}$ where the shorthand notations refer to the following:

D: The agent *does not forward* its packet to the other network (i.e. agent chooses the non-cooperative route) and *drops* all packets from other network if asked for help to forward the packets.

F: The agent *forwards* its packet to the other network (i.e. agent chooses the cooperative route) and in turn forwards all packets if the other network asked for help to forward the packets.

Each of the joint actions incurs some cost and payoff associated to it. In this chapter, we propose a payoff function according to the payoff matrix in Table 3.1.

Table 3.1: Payoff matrix of interaction between sensor nodes in different domains

	= D	**= F**
= D		$0, -$
= F	$-$, 0	$($ $-$ $), ($ $-$ $)$

The physical interpretation is as follows. Suppose that both domains choose the action *{D,D}* where each domain denies cooperating with the other domain

to forward the other domain's packets, the there is no energy cost in helping the other domain. The only payoff is the packet received rate (PRR), μ_i, for domain $i = 1,2$ which is the average ratio of correctly received packets at a sink in such path. For a transmission of a data packet of length *b* bits, PRR can be expressed as

$$PRR = (1 - P_b)^b , \tag{3.1}$$

where P_b is the bit error probability for one hop communication using OQPSK modulation for Zigbee devices operating at 2.4GHz (Ahmed and Fisal, 2008). The PRR ranges between 0-1 and reflects the benefits in terms of the reliability of the route. If either agents decides to help the other domain forward its packets, while the other domain declines to cooperate, the action would be {D,F} or {F,D}. In this case, the cost of cooperation would be the energy the domain *i* uses to help forward the (domain -*i*) packets to the sink ($-\eta_i$). The minus sign depicts the energy consumed perceived as a cost of cooperation. The other domain is zero since it does not incur cost (as it refuses to cooperate) but does not enjoy any benefits of increased PRR. If both agents agree to cooperate, the action would be {F,F} and the associated payoff is the PRR (μ_i) and the net gain in energy savings. The energy saving is determined from the energy consumption on Non-cooperative path (γ_i) subtracted by the energy consumption on cooperative path (η_i). Note that a positive net energy gain ($\gamma_i - \eta_j$) from cooperation and high PRR will result in both agents selecting the cooperative action {F,F}.

The decision to select a joint action in this Non-cooperative game depends on the NE with respect to the payoff matrix in Table 3.1. A joint action of

the stage game is Nash equilibrium point if every agent receives its highest payoff at this point. Let (a_1, a_2) be a joint action of agents in both domains where $a_i \in A_i$. A joint action (a_1^*, a_2^*) is said to be a NE if agent *i* selecting action $a_i \in A_i$ gives the highest payoff when its opponent selects its best action a_{-i}^* where $a_{-i} \in A_{-i}$ is an action of the agent i's opponent $(-i)$. The NE can be presented as

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*) \tag{3.2}$$

where $u_i(a_i, a_{-i}^*)$ is the value of the payoff function under joint action (a_i, a_{-i}^*) of agent *i*.

Typically NE corresponds to a pure-strategy equilibrium, which is a condition that an agent can choose with certainty an action which achieves the best response to the other agent's choice. However, not all games have pure-strategy equilibrium. Under this circumstance, the agents need to select their actions randomly according to some probability to achieve the NE. Such equilibrium is called a mixed-strategy equilibrium. The Lemke Howson method (LH), is the best known method to solve for mixed-strategy NE between two agents (Shoham and Brown, 2009). The advantage of LH method is that it is guaranteed to find at least one NE point. In this thesis, the LH method is therefore used in the Non-cooperative game when there are multiple NE or when pure strategy NE does not exist. The pseudo code of NCG-LH is shown in Figure 3.2.

3.3.4 Compared algorithms

In order to evaluate the performance of Non-cooperative game routing using LH algorithm (NCG-LH) for packet forwarding in multi-domain WSNs, we compared it with a variations of the AODV routing protocol which is used in IEEE

standard 802.15.4 ZigBee protocol stack. In particular, there are two variations of the AODV scheme compared in this experiment: is the Non-cooperative AODV routing (No cooperation routing) and the cooperative AODV routing (All cooperation routing). The Non-cooperative AODV routing uses AODV to discover the least energy consumption route consisting of nodes within the same domain. On the other hand, the cooperative AODV routing discovers the least energy consumption route which consist of nodes from the other domain.

```
BEGIN
  for topology 1:100
    Initialize energy for each node to full battery level
    Let t=0
      do
        Random source node to create data packet
        Establish two routing tables using AODV routing protocol
        (one table for paths in own network and another one for paths in cooperative networks)
        Calculate payoff value for all available action following Table 3.1
        Determine strategy using NE and LH method
        Sent data packet to sink following its strategy

        Let t=t+1
      while (at least one node run out of battery )
   endfor

END
```

Figure 3.2 Pseudo code of NCH-LH algorithm

3.4 Experiment results

In this section, we evaluate the performance of the proposed NCG-LH algorithm and investigate the cooperative conditions of the packet forwarding

strategies in multi-domain WSNs. We study its performance under the uniform random topology and tree topology models.

We consider two WSNs co-existing in the same area, which are deployed in a multi-domain WSN. In each WSN, the source nodes act as individual agents, which make their own decisions which route to forward their packets to. Then all immediate nodes belonging the chosen route act according to the source nodes decision. The goal of each agent is to maximize the packet delivery within its network to the sink by based on the energy payoff matrix in Table 3.1. We investigate two scenarios, uniform random topology and tree topology scenarios. The purpose of studying the random topology scenario is to investigate how the unguaranteed connectivity between sensor nodes and the common sink affects cooperation. On the other hand, the tree topology scenario is studied to study the effect of guaranteed connectivity on node cooperation. Simulation results are carried out over 100 randomly generated topologies to avoid performance bias based on a particular topology.

To study the effect of cooperation between nodes in multi-domain WSNs, the following performance metrics are measured:

1) Cooperation: the ratio of the number of routes using nodes from both domains to the total number of routes discovered.

2) Packet delivery ratio (PDR): the ratio of the number of data packets received over the number of data packets sent out.

3) Network lifetime: the time at which the first network node runs out of energy.

4) Fairness: the difference in average energy consumed along a forwarding path in domain 1 and 2 is used to evaluate the fairness of the algorithms. The discrepancy in energy usage between the two domains indicates some degree of unfair resource

allocation between the domains as one domain is utilized more than the other. Therefore, fairness is achieved when this discrepancy is reduced to zero indicating that energy in both networks are equally used.

The metrics compared are averaged from the measurements obtained from both networks.

3.4.1 Uniform random topology

We consider two WSNs co-existing in the same area, which are deployed in a 500x500 m^2 in multi-domain WSNs as shown in Figure 3.3. The simulation parameters are shown in Table 3.2.

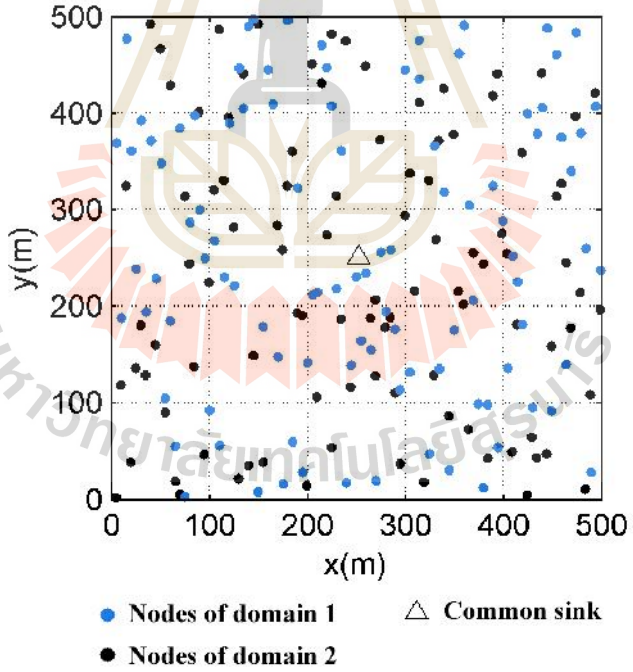


Figure 3.3 Uniform random topology for 100 nodes per domain

Table 3.2: Parameter setting for uniform random topology

Parameter	Value
Number of domains	2
Number of sensors per domain	20, 40, 60, 80, 100
Area size	500x500 m^2
Sink position	(250,250)
Number of maximum hop	5 hops
Transmission range	100 m
Data load per packet, *b*	100 *bytes*
Path loss exponent, σ	2, 4
Number of failed nodes	12-48
Routing protocol	AODV routing
Distribution of the sensors	Uniform random topology
Random Topology	100

3.4.1.1 Effect of density

In order to identify the parameters which affect cooperation in multi-domain WSNs, we varied the node density by increasing the number of sensors in each domain as well as network connectivity (i.e. unguaranteed and guaranteed connectivity)

Figure 3.4-3.7 presents the performance comparison of NCG-LH routing algorithm at different node density under uniform random topology. Figure 3.4 shows the average proportion of cooperation with varying number of sensor nodes per domain which represents the density of each network. In this figure, results of only NCG-LH algorithms because All cooperation (which only uses routes consisting of nodes from both domains) and No cooperation (which only uses routes consisting of nodes from the same domain) always have a proportion of cooperation

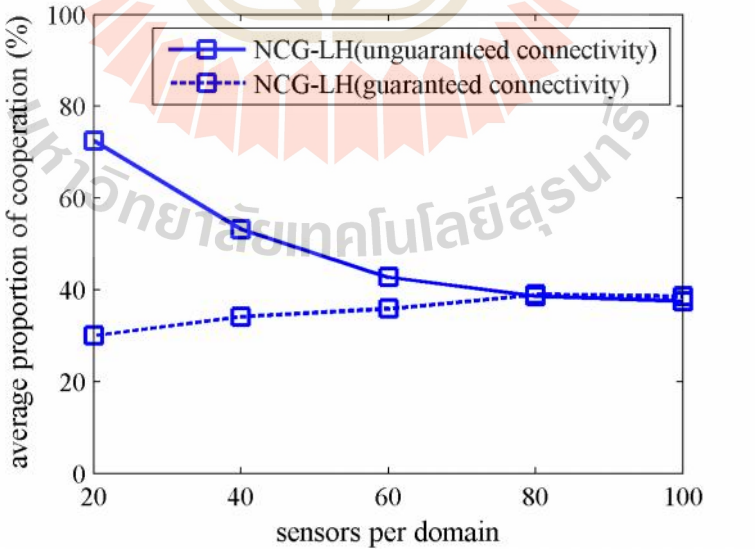


Figure 3.4 Average proportion of cooperation at different node density under uniform random topology

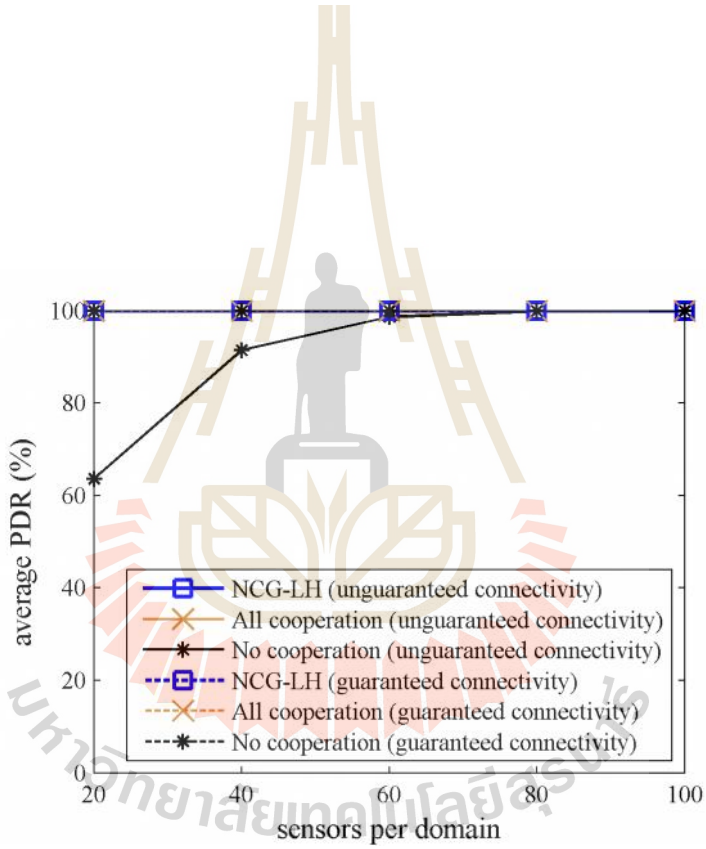


Figure 3.5 Average packet delivery ratio at different node density under uniform random topology

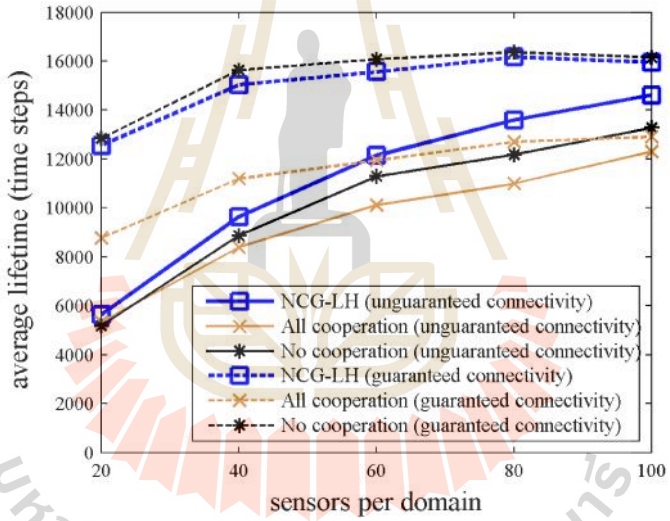


Figure 3.6 Average network lifetime at different node density under uniform random topology

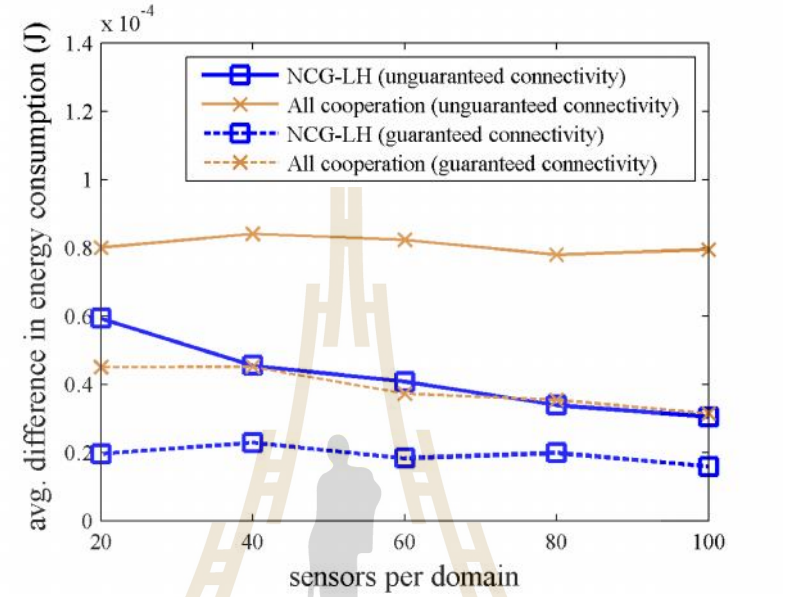


Figure 3.7 Average difference in energy consumption at different node density under uniform random topology

Figure 3.8-3.11 shows the performance comparison of NCG-LH algorithms under failure prone and hostile environments for uniform random topology. Figure 3.8 depicts the proportion of cooperation. The results show that the proportion of cooperation by NCG-LH algorithm is higher when node failure and path loss exponent is increased. This suggests that cooperation is imperative in harsh environments.

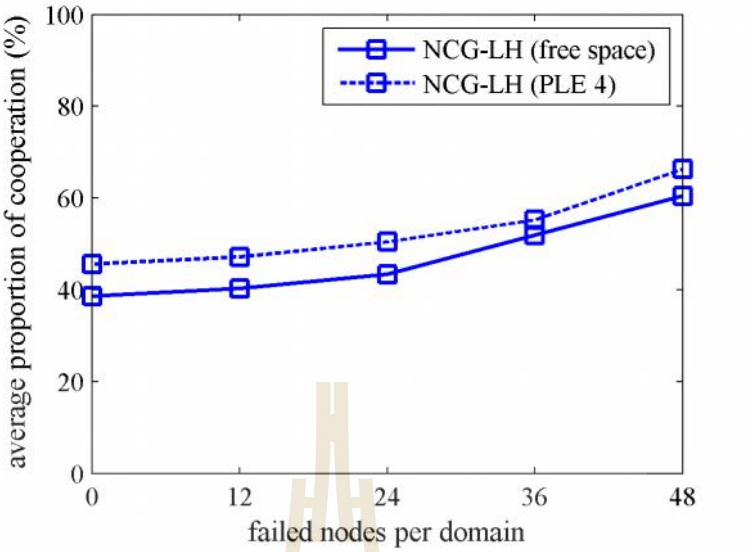


Figure 3.8 Average proportion of cooperation in hostile environment under uniform random topology

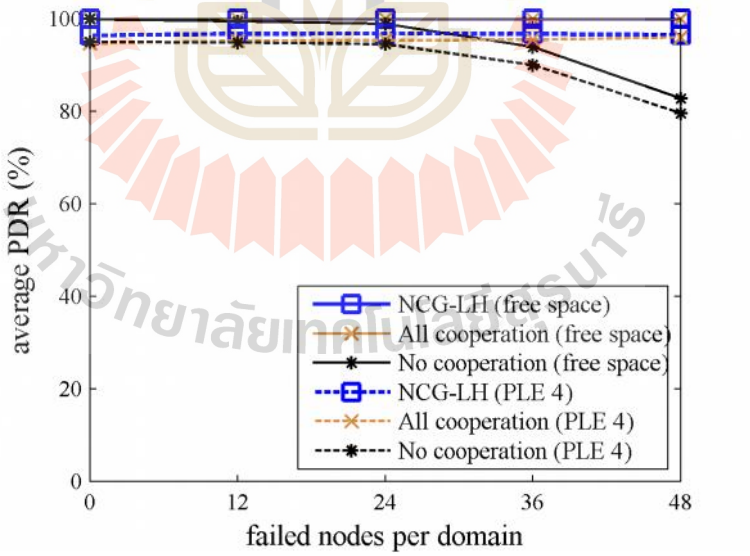


Figure 3.9 Average packet delivery ratio in hostile environment under uniform random topology

Figure 3.9 shows the average successful packet delivery ratio. In case of free space with 0-12 failed nodes, the figure shows that the PDR of all three algorithms are comparable. However, when the number of node failures increases to 24, 36 and 48, the PDR of No cooperation algorithm gradually drops to 80%. On the other hand, NCG-LH and All cooperation algorithms can maintain their PDR at 100%. The reason is because the cooperation between the two domains which permits alternative routes to avoid the failed nodes. This clearly shows that cooperation is necessary when network is prone to node failure. As the path loss exponent increases, the figure demonstrates that the proposed NCG-LH algorithm can perform as well as All cooperation algorithm by maintaining PDR at 90% on average. Moreover, NCG-LH algorithm outperforms No cooperation algorithm by over 24% PDR when the path loss exponent is 4. Hence, No cooperation cannot maintain acceptable PDR in the presence of failed nodes and higher PLE. On the other hand, NCG-LH cannot only maintain high PDR but also in an energy efficient manner as illustrated by the longer network lifetime than All cooperation algorithm in Fig 3.10. In particular, NCG-LH can attain an average of 14.8% and 22.5% longer network lifetime than All cooperation algorithm at PLE 2 (free space) and 4, respectively.

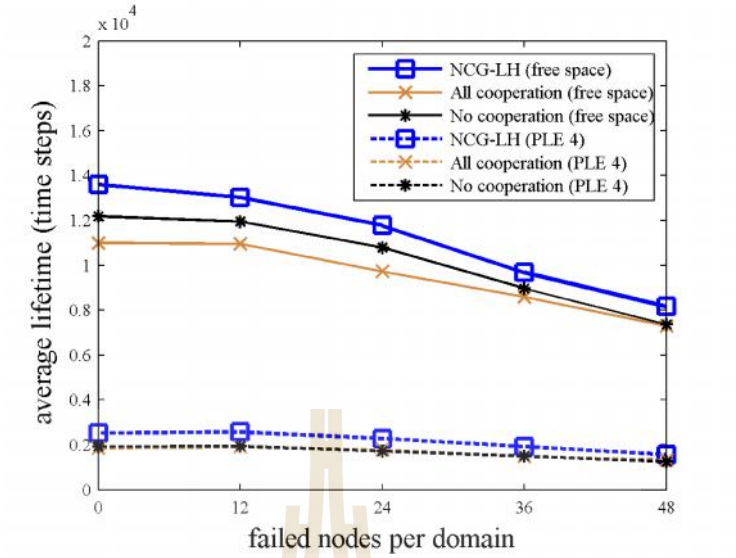


Figure 3.10 Average network lifetime in hostile environment under uniform random topology

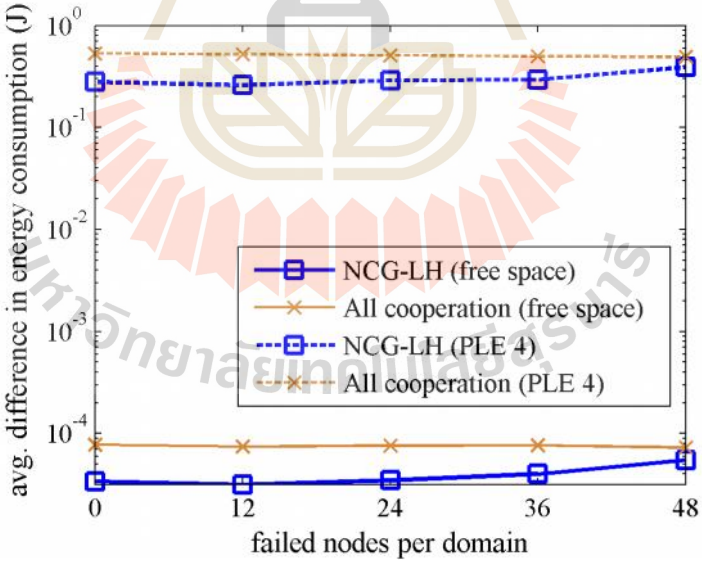


Figure 3.11 Average difference in energy consumption in hostile environment under uniform random topology

Figure 3.11 show that NCG-LH can also attain more fairness by attaining lower difference in energy consumption between the two domains than All cooperation algorithm. Moreover, it can be seen that in case of PLE 4 there is more difference in energy consumption than in the free space case. This because PLE 4 case consumes more energy than in the free space case as seen in the transmission cost of agent, $E^{j}_{i,TX}(b,d) = E_{elec} \times b + (\pounds_{amp} \times b \times d^{\pm})$, where $\pm$ is path loss exponent.

The results obtained from the uniform random topology suggest that cooperation in multi-domain WSNs are not always be beneficial to any network and may even waste energy and reduce network lifetime. Cooperation between networks are beneficial if 1) the networks are sparse and have no guaranteed connectivity; 2) the networks is prone to faulty nodes which may cause disconnected routes; 3) in presence of hostile network environment (high path loss). In such scenarios where cooperation is required, NCG-LH has shown to select suitable actions giving rise to high PDR and longer network lifetime than algorithms which always cooperate (All cooperation) and do not cooperate at all (No cooperation).

3.4.2 Tree topology

In this section, we present results of another realistic topology of WSN deployment for many applications in recent years. In the previous section, the uniform random topology was more suitable for random deployments such as scattered sensor placements in a large area (e.g., forests, farm land). There are other applications which required structured tree topologies which also ensure guaranteed connectivity in large coverage areas (Guizani et al., 2015).

Therefore this section, the tree topology for multi-domain WSNs is investigated. We consider two WSNs co-existing in the same area, which are deployed in a 3000x3000 m^2 as shown in Figure 3.12. The simulation parameters are shown in Table 3.3. Because tree topology requires guaranteed connectivity, we therefore consider only this scenario in this section.

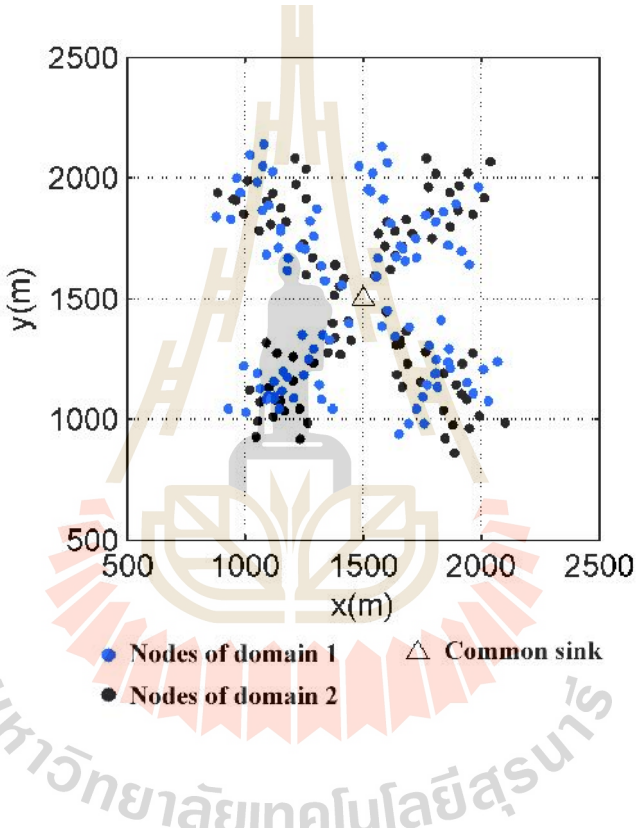


Figure 3.12 Tree topology for 100 nodes per domain

Table 3.3: Parameter setting for tree topology

Parameter	Value
Number of domains	2
Number of sensors per domain	20, 40, 60, 80, 100
Area size	3000x3000 m^2
Sink position	(1500,1500)
Number of maximum hop	10 hops
Transmission range	100 m
Data load per packet, *b*	100 *bytes*
Path loss exponent, σ	2, 4
Number of failed nodes	1-4
Routing protocol	AODV routing
Distribution of the sensors	Tree topology
Random Topology	100

3.4.2.1 Effect of density

The performance of all algorithms in guaranteed connectivity scenario at different node density under tree topology are shown in Figure 3.13-3.16. Figure 3.13 shows that the proportion of cooperation obtained from NCG-LH remains between 40-45% as the network density increases suggesting that node density has little effect on cooperation when connectivity to the sink is guaranteed. The observed cooperation between nodes is due to the energy savings only, not from the presence of node density in the network.

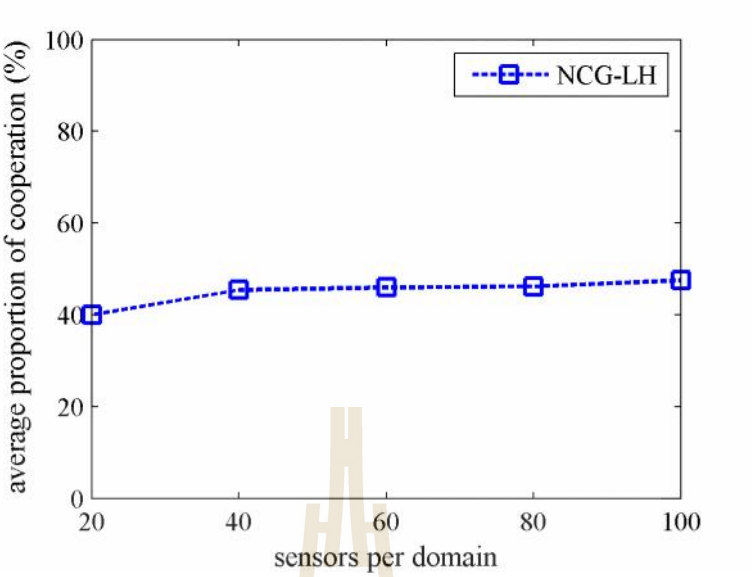


Figure 3.13 Average proportion of cooperation at different node density under tree topology

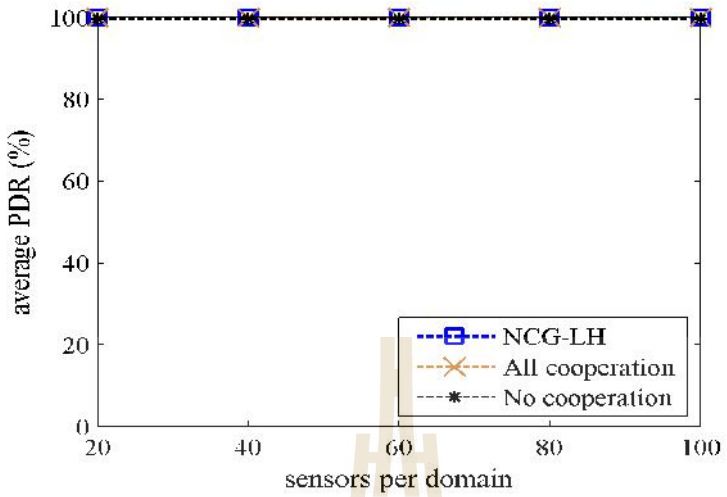


Figure 3.14 Average packet delivery ratio at different node density under tree topology

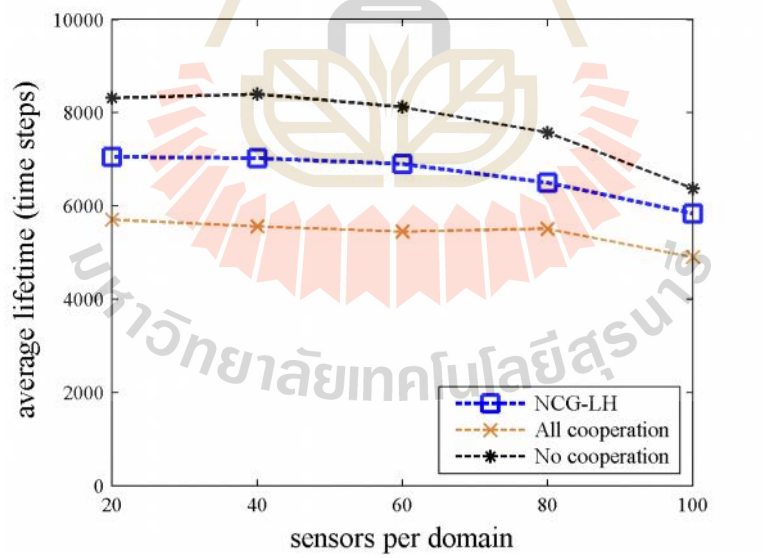


Figure 3.15 Average network lifetime at different node density under tree topology

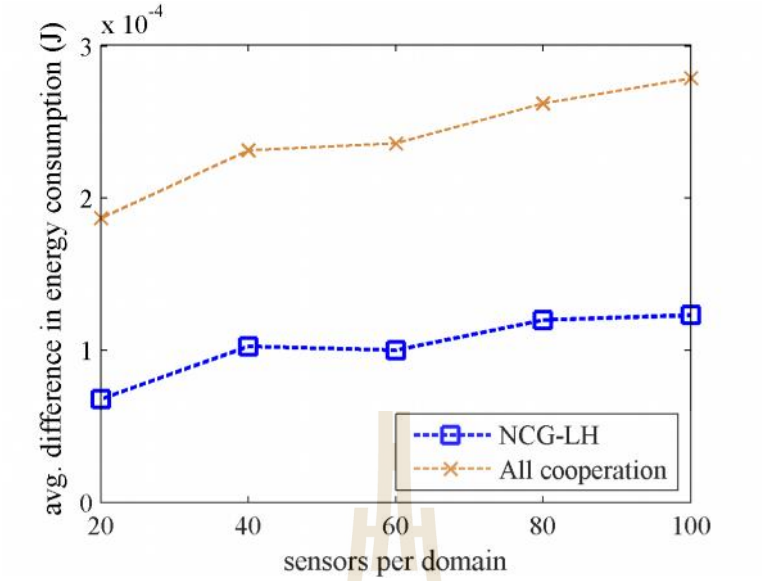


Figure 3.16 Average difference in energy consumption at different node density under tree topology

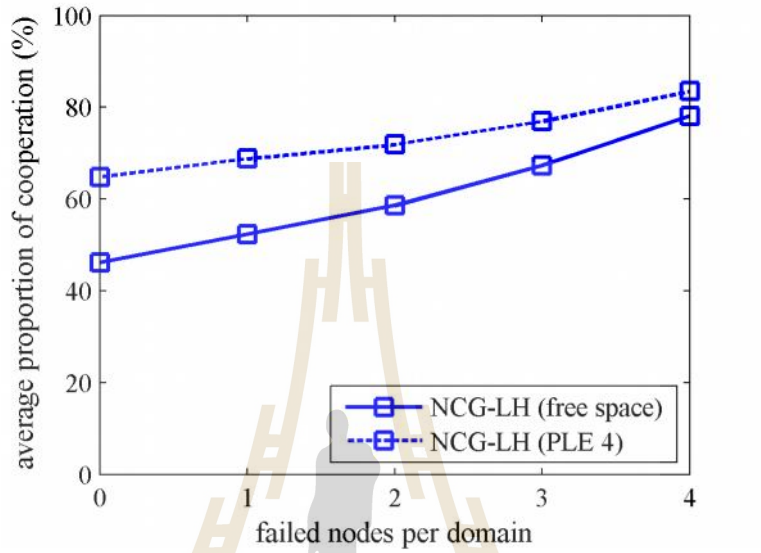


Figure 3.17 Average proportion of cooperation at hostile environment under tree topology

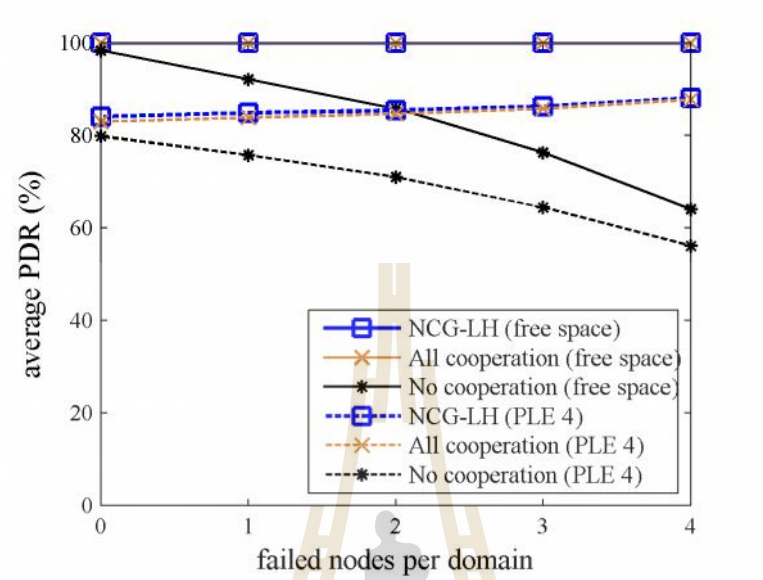


Figure 3.18 Average packet delivery ratio at hostile environment under tree topology

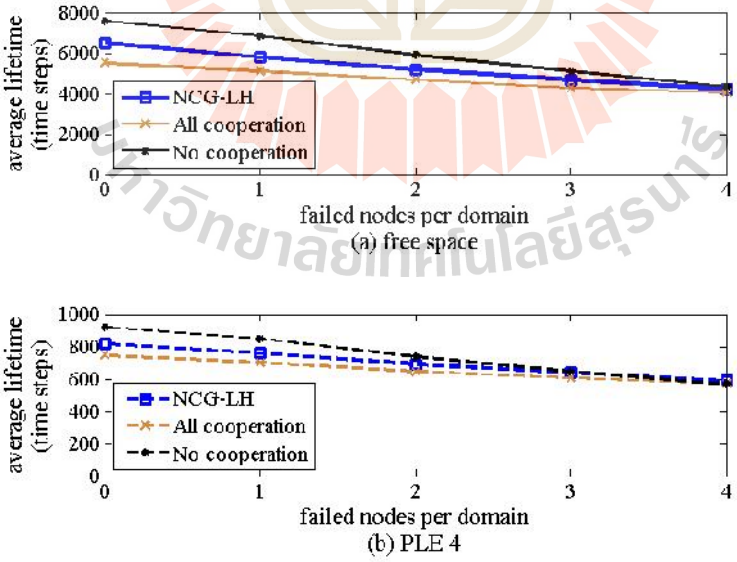


Figure 3.19 Average network lifetime at hostile environment under tree topology

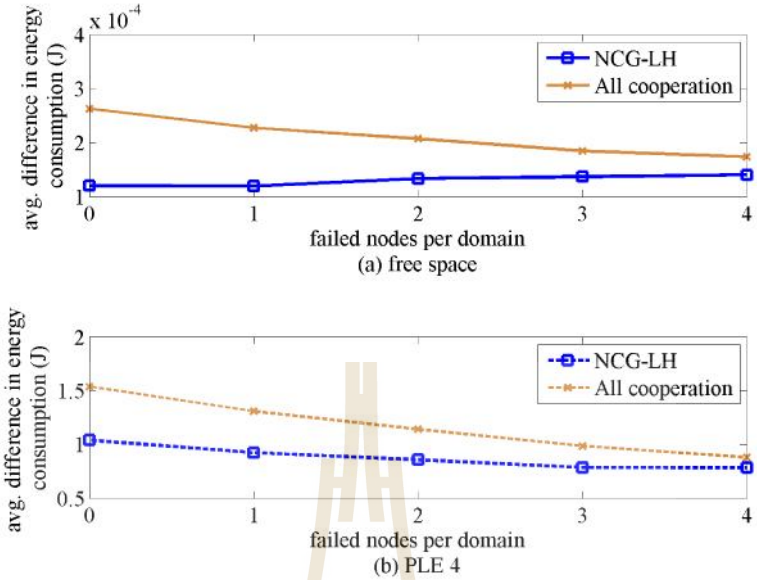


Figure 3.20 Average difference in energy consumption at hostile environment under tree topology

However, in harsher environments, cooperation becomes a necessity and NCG-LH selects its actions to increase cooperation among nodes.

3.5 Summary

In this chapter, we apply the distributed Non-cooperative game algorithm based on Lemke Howson method to the packet forwarding game for a multi-domain WSN to determine the best actions to attain mutual benefits for both networks. The contribution of this chapter is three-fold; 1) we show that NCG-LH algorithm can be applied to achieve the best mutual packet forwarding policy in Non-cooperative multi-domain WSNs in a distributed manner; 2), we evaluate NCG-LH algorithm and identify parameters that affect cooperation between networks and fairness of benefits that the networks can achieve; 3) we propose a novel payoff matrix for packet forwarding game in Non-cooperative multi-domain WSNs.

Results have been evaluated in both the uniform random (unguaranteed and guaranteed connectivity) and tree topologies (guaranteed connectivity), under varied node density, failed nodes and path loss exponents. It was found that cooperation was not always necessarily beneficial to all networks. In particular cooperation is beneficial is required when 1) the network is sparse and connectivity is not guaranteed; 2) the network is prone to failed sensor nodes which affect reliability of the routes; 3) the network is deployed in harsh environments with path high loss. Under such scenarios, cooperation among nodes permits use of diverse routes which enhance reliability and prolong network lifetime for all networks. On the other hand, in networks with guaranteed connectivity, results suggest that cooperation is unnecessary and can result in shortened network lifetime.

Results show that the NCG-LH algorithm is able to select appropriate actions to forward the packets. NCG-LH action selection has the adaptability to various network configuration (network connectivity, node density, failed nodes and path loss exponent) resulting in high PDR, prolonged network lifetime and fair energy consumption among the domains. This is due to the selection of NE and the Lemke Howson method in the NCG-LH framework based on the payoff matrix which takes into consideration the benefits of all domains.

CHAPTER IV

FAIR ROUTE SELECTION IN MULTI-DOMAIN WSNS USING NON-COOPERATIVE GAME THEORY UNDER SEPARATE SINK SCENARIO

4.1 Introduction

In recent years, multiple WSNs have been constructed within the same interesting region (Fadel et al., 2015; Rashid et al., 2016). For such cases, researchers have been investigating cooperation among sensor nodes belonging different network authorities which could potentially gain certain benefits. Such benefits include alternative routing paths and reduced energy consumption, which can prolong their network lifetime and enhance reliability of packet delivery. Some routing protocols for multi-domain WSNs have been proposed under common sink scenario (Felegyhazi et al., 2005;Wu et al., 2005; Vaz et al., 2008; Singsanga et al., 2010). The sink node is shared by the multiple networks, and located at the center of the area of interest. The previous chapter introduced the application of non-cooperative game theory to address cooperation problem in multi-domain WSNs and proposed a routing algorithm named *Non-cooperative game algorithm based on Lemke Howson method* (NCG-LH) algorithm. The performance of the proposed algorithm was evaluated in a common sink scenario in order to conceptually show that non-cooperative game theory can be applied to solve the non-cooperative packet forwarding problem in

distributed multi-domain WSNs. The algorithm was shown to determine a suitable packet forwarding strategy between multiple domains that can extend the network lifetime and enhance reliability by using Nash equilibrium (NE) and Lemke Howson method. However, in the previous chapter, the common sink scenario was investigated to analysis the solution. In real world WSN applications, each network normally has its own sink. There are several recent researches solve routing problems in multi-domain WSNs under separate sink scenario for more realistic formulation (Yang et al., 2007; Bicakci et al., 2013; Rovcanin et al., 2014; Singhanat et al., 2015; Kinoshita et al., 2016). Therefore, to evaluate the proposed NCG-LH algorithm in a more realistic sink scenario is needed.

In this chapter, NCG-LH algorithm in Chapter 3 is evaluated in a multi-domain WSN with separate sink scenario. Similar to Chapter 3, the parameters that effect cooperation between multiple co-located WSNs are also studied in this chapter i.e. network density, node failure, path loss exponent and network connectivity. This chapter additionally investigated the other parameters that effect cooperation i.e. the difference in node density in each domain and sink positions. The performance is compared with 3 existing algorithms including variations of the AODV routing protocol i.e. 1) the AODV routing with no cooperation, 2) the cooperative AODV routing 3) an existing algorithm called pool-based routing algorithm (Kinoshita et al., 2016). While the first two algorithms was adopted from Chapter 3, the last algorithm takes into account of fair route selection in multi-domain WSNs. The simulation results are evaluated in uniform random topology only. This is because the proposed algorithm can distinctly provide the best performance in uniform random topology. The environment setting, configuration and network model are all the same as

Chapter 3 except number of sinks and their positions. The results show that by using the proposed algorithm which provides fair route selection, all networks can gain longer network lifetime.

The main contributions of this chapter are three-fold: 1) The non-cooperative game algorithm (NCG-LH) is applied to a non-cooperative multi-domain WSN under a separate sink scenario; 2) Investigation of fairness in terms of the difference in energy consumption between domains and comparison between a game theoretic approach (NCG-LH); and non-game theoretic technique (Pool-based method); 3) Identification of parameters that effect cooperation between multiple co-located networks and fairness.

4.2 Simulation results

In this section, we provided the simulation results of the proposed NCG-LH algorithm performed in Visual C++ environment and investigate the cooperative conditions of the packet forwarding strategies in multi-domain WSNs under separate sink scenario. We consider two different WSNs, N_i, $i=1,2$, co-existing in a multi-domain WSN. Each WSN domain deployed randomly v sensor nodes, $N_i = \{n_i^1, n_i^2, \ldots, n_i^v\}$, and one sink. The simulation environment is set to be a square area of 2500 m^2. In each time step, each WSN chooses randomly a source node to send data packet to its sink. Source node acts as an agent of the packet forwarding game to determine a fair routing policy by using Nash equilibriums (NE) to achieve a policy in order to prolong the network lifetime by using the proposed algorithm.

Similarly to Chapter 3, simulations in this chapter are then carried out under varying number of nodes, number of failed node and path loss exponent as well as

network connectivity (i.e. unguaranteed and guaranteed connectivity). This is because these factors can cause connectivity problems which create failure in the forwarding path thereby reducing the reliability of WSN. Under such scenarios, NCG-LH exhibits the ability to allocate resource sharing between multiple networks and determine a fair routing for packet forwarding to eliminate connectivity weakness and prolong network lifetime.

We compare the proposed NCG-LH algorithm with 3 existing algorithms in 3 metrics including:

- *Proportion of cooperation:* the ratio of the number of cooperative routes to the total number of routes discovered.
- *Network lifetime:* The lifetime of each network. Since each time step, a packet is transmitted, this thesis thus measures the network lifetime in terms of the total number of time steps that data packet transmitted at the sink node until the first node dies.
- *Fairness:* the difference in average energy consumed along a forwarding path between network domain N_1 and N_2. From a fairness point-of-view, energy in different network domain should be consumed equally. If one domain uses more energy than the other domain, there will be a discrepancy in energy consumption between domain 1 and domain 2.

The simulation results are divided into 3 scenarios. The simulation parameters are shown in Table 4.1. The other environments and configurations are similar to Chapter 3. Simulation results are carried out over 100 randomly topologies. The

experimental results are shown in this section obtained from average results from both domains.

Table 4.1: Parameter Setting

Parameter	**Value**		
	Scenario 1	**Scenario 2**	**Scenario 3**
Area size	500x500 m^2		
Number of domains	2		
Number of sensors per domain	20 - 100	80- 240	80- 240
Sink position of P_{S_1}	[125,250]	[125,250]	[0,0]
Sink position of P_{S_2}	[375,250]	[375,250]	[500,500]
Distribution of the sensors	Uniform random		
Number of maximum hop	5 hops		
Transmission range	100 m		
Data load per packet, *b*	100 *bytes*		
Path loss exponent, σ	2, 4		
Number of failed nodes	4-48		
Routing protocol	AODV routing		

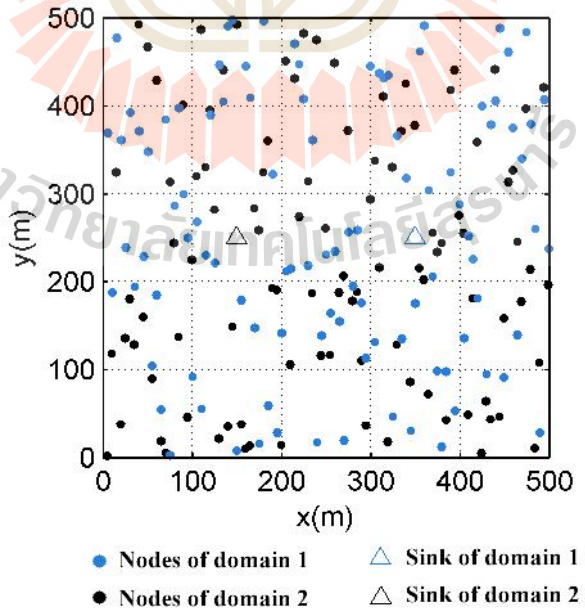


Figure 4.1 Uniform random topology for 100 nodes per domain

. Ea

on at N_1 a

N_2. S

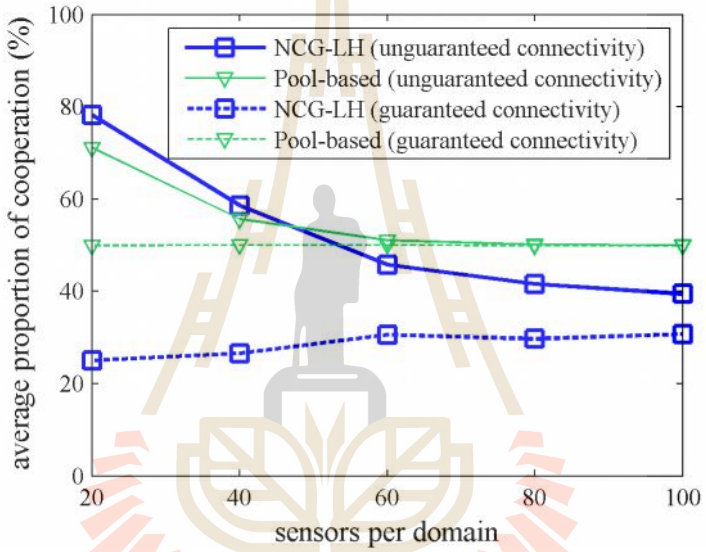
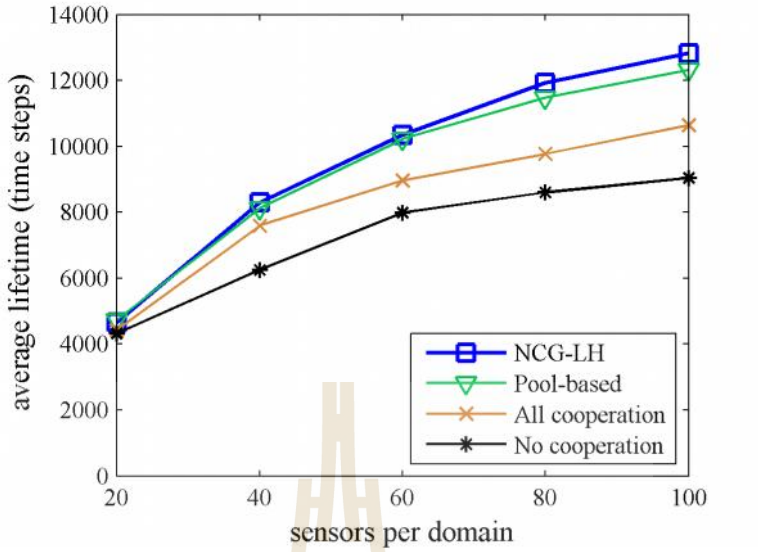
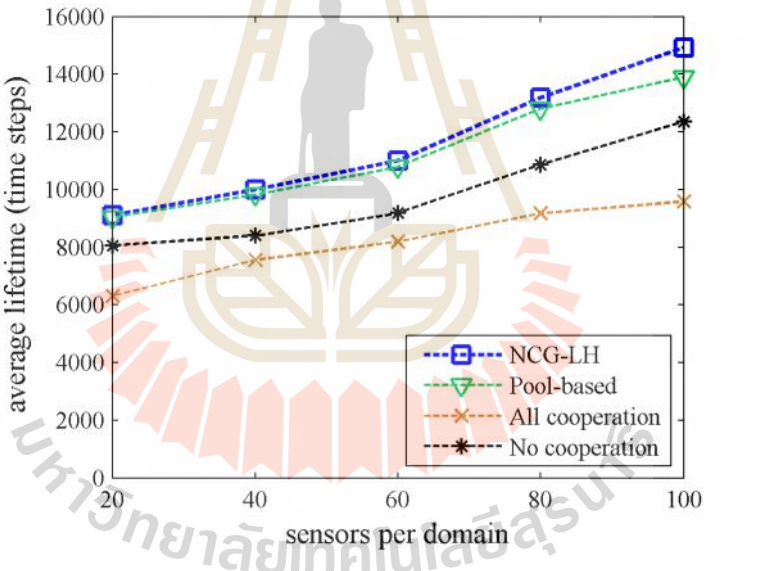


Figure 4.2 Average proportion of cooperation at different node density

that when the node density is increased, NCG-LH tends to demote cooperation between two different domains from 80% to 40%. It is because the higher the node density, more paths will be available for sensors to send packets to the sink. So cooperation between both agents is not necessary. This suggests that cooperation is required if the density of sensors is low. Moreover, it can be seen that Pool-based algorithm is comparable to NCG-LH when network density is low. But when network density is enough to provide multiple paths to send packets to the sink, the proportion of cooperation from Pool-based is always 50%. It is because Pool-based algorithm always balances the load between cooperative path and path with no cooperation. In the case of guaranteed connectivity, the figure shows that the proportion of cooperation of both NCG-LH and Pool-based routing algorithms are almost constant as node density increases. This because each network has high connectivity, thus the cooperation is not required. However, NCG-LH requires 15% less proportion of cooperation than Pool-based algorithm on average as node density increases to 100 nodes per domain.



(a) unguaranteed connectivity



(b) guaranteed connectivity

Figure 4.3 Average network lifetime at different node density

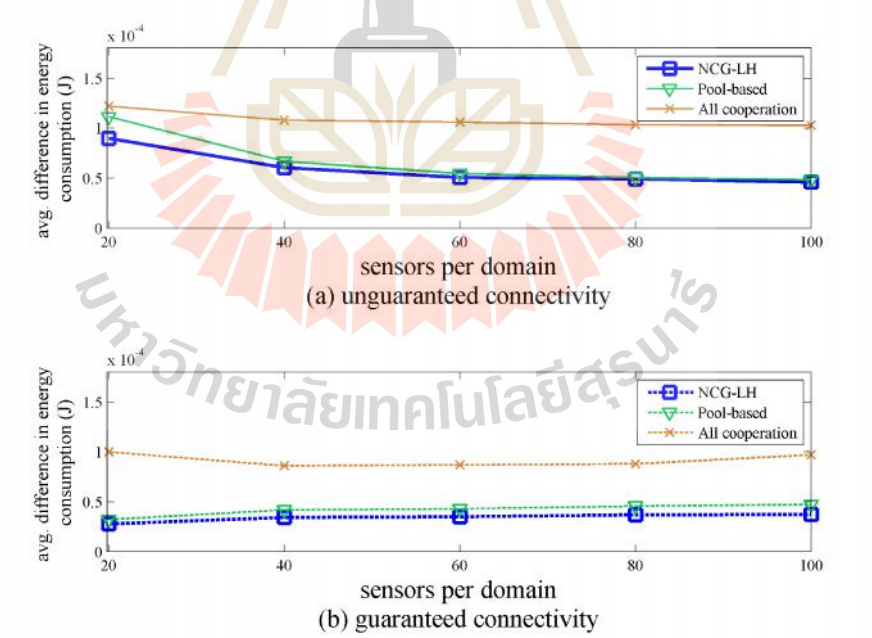


Figure 4.4 Average difference in energy consumption at different node density

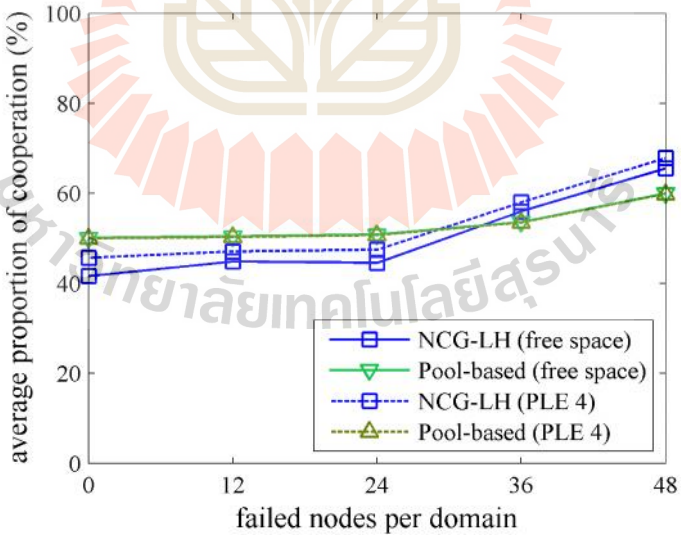
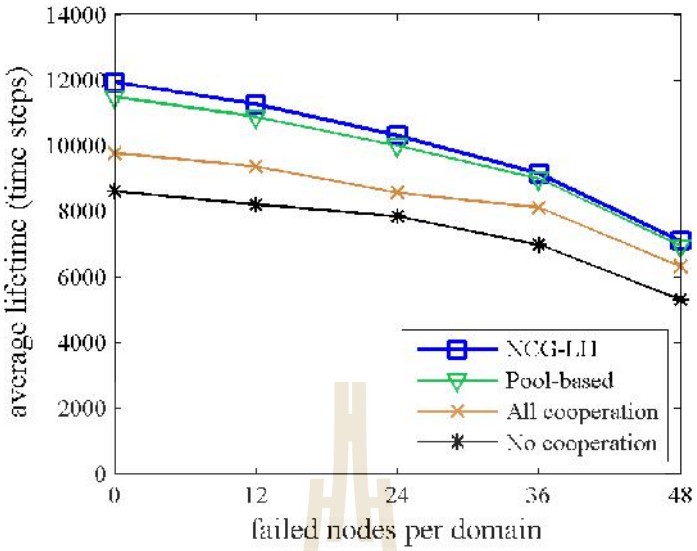


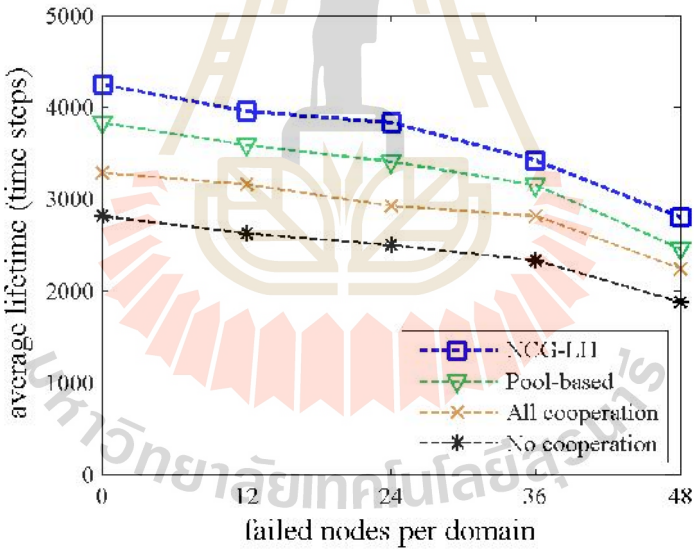
Figure 4.5 Average proportion of cooperation in various node failures under different path loss exponents

Figure 4.5 shows the average proportion of cooperation by varying the number of failed nodes per domain with PLE in the range of 2 to 4. The figure depicts that both algorithms provide more cooperative strategy when number of failed nodes and PLE increase in order to avoid disconnectivity. However, NCG-LH can prolong network lifetime than Pool-based algorithm as shown in Figure 4.6. In free space, NCG-LH achieves longer network lifetime than Pool-based, All cooperation and No cooperation routing algorithms by 4.3%, 16% and 25.5%, respectively, on average as the number of failed nodes increases. With PLE 4, a similar trend is found in free space with NCG-LH obtaining longer network lifetime than Pool-based, All cooperation and No cooperation routing algorithms by 9%, 19.6% and 31.2%, respectively, on average as the number of failed nodes increases. Note that, in case of PLE 4, NCG-LH attains longer network lifetime than free space case when compared with existing algorithms. This because NCH-LH takes the path loss exponent parameter into account in the calculation of energy consumption, then chooses the action with the maximum energy saves whereas the other existing algorithms do not. NCG-LH therefore chooses suitable actions that can prolong network lifetime better than other algorithms.

Figure 4.7 shows average fairness in energy consumption with a varying number of failed nodes under different PLE. It can be seen that, even though NCG-LH and Pool-based algorithms provide different fair packet forwarding policy, NCG-LH can provide an average energy consumption close to that of Pool-based algorithm. Moreover, NCG-LH also achieves low discrepancy in energy consumption compared to All cooperation routing algorithm. This suggest that sharing resources is not always a fair strategy for both networks.



(a) free space



(b) PLE 4

Figure 4.6 Average network lifetime in various node failures under different path loss exponents

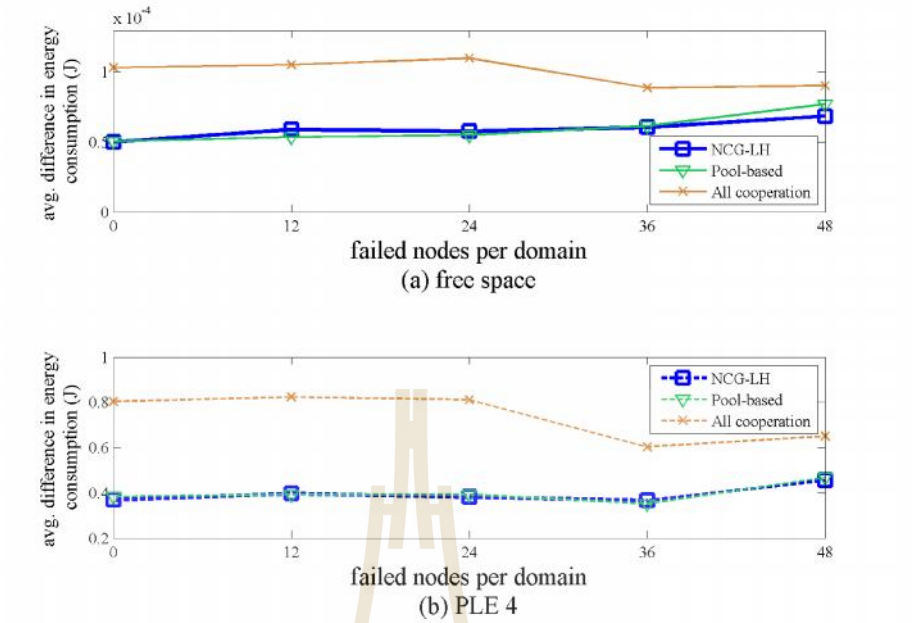


Figure 4.7 Average difference in energy consumption in various node failures under different path loss exponents

sity

In th N_1 is

N_2 is

N_1 N_2

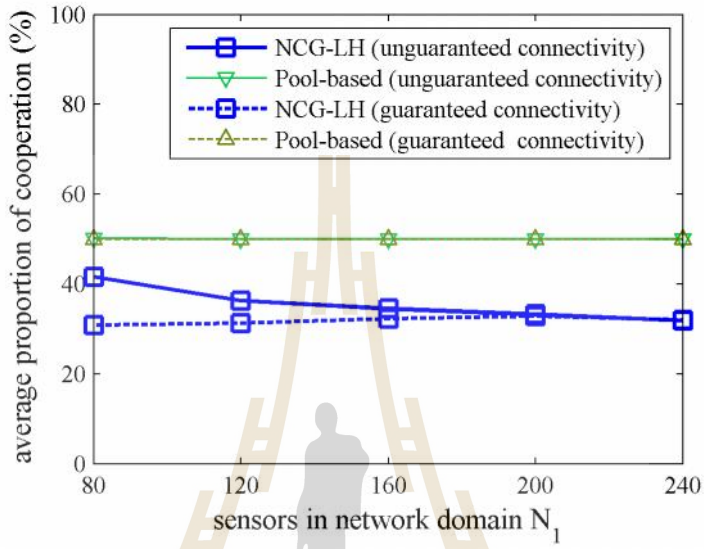


Figure 4.8 Average proportion of cooperation at different node density of network N_1

show

N_1. I

of c

N_1 in

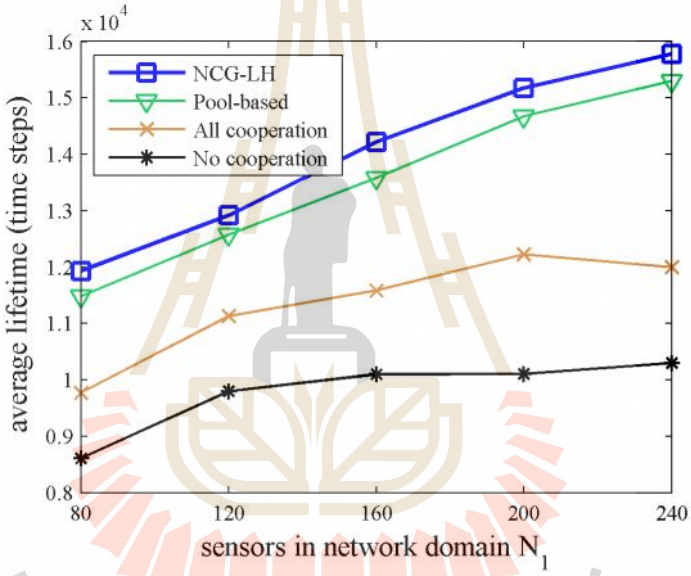
t low coop

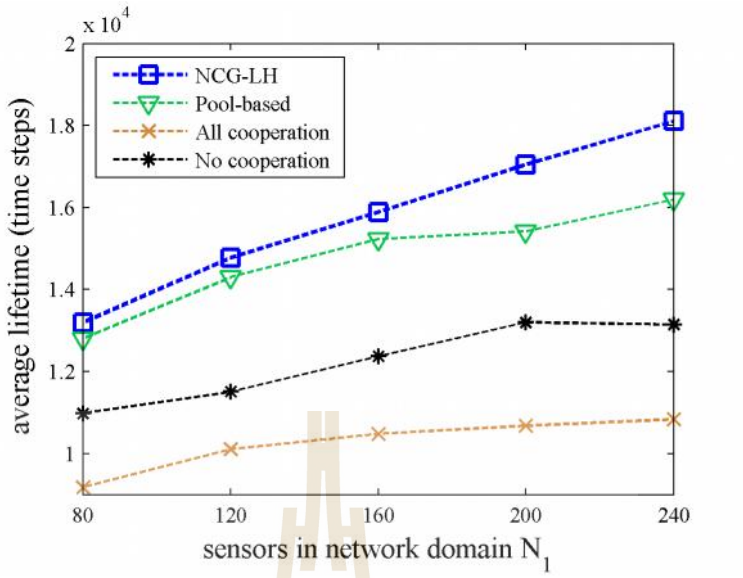
N_1 a N_2 h

N_1

N_2.

N_1





(b) guaranteed connectivity

Figure 4.9 Average network lifetime at different node density of network N_1

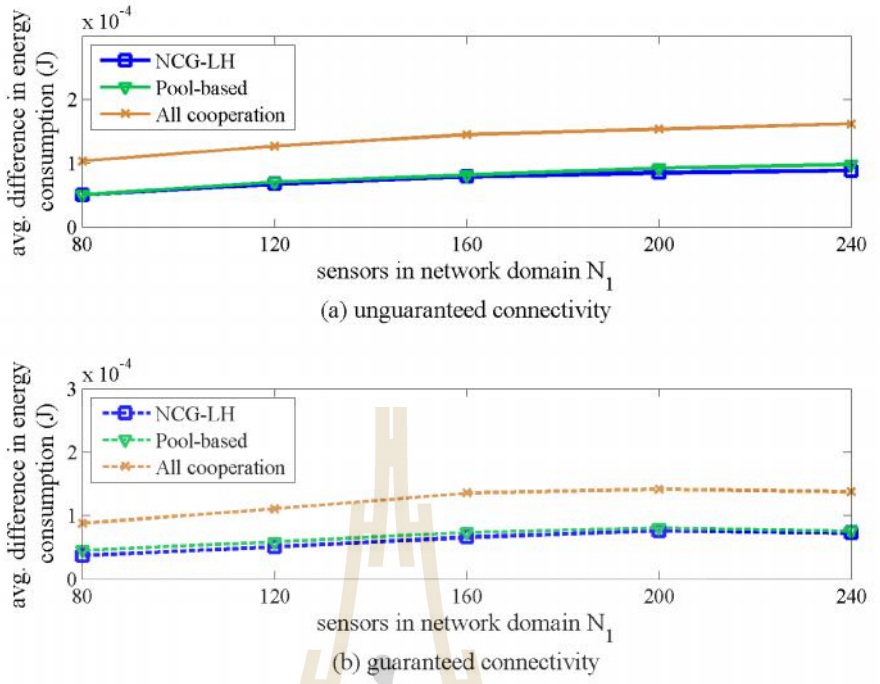


Figure 4.10 Average difference in energy consumption at different node density of network N_1

own i ducte

N_1 at N_2 at

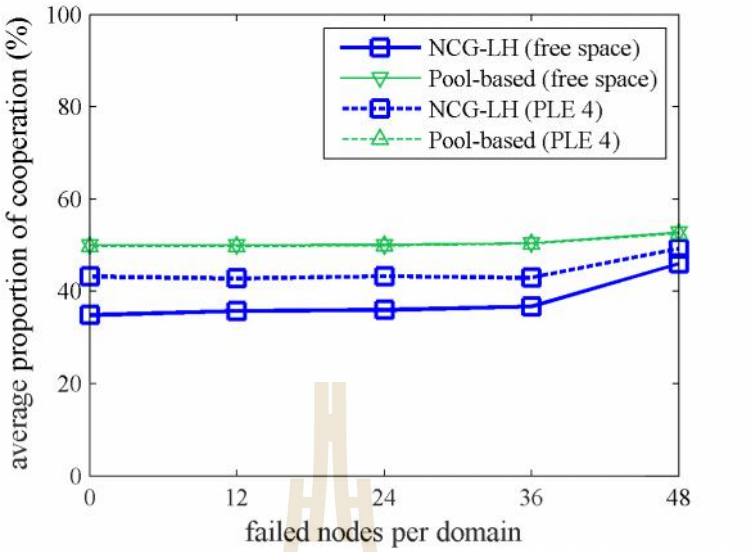
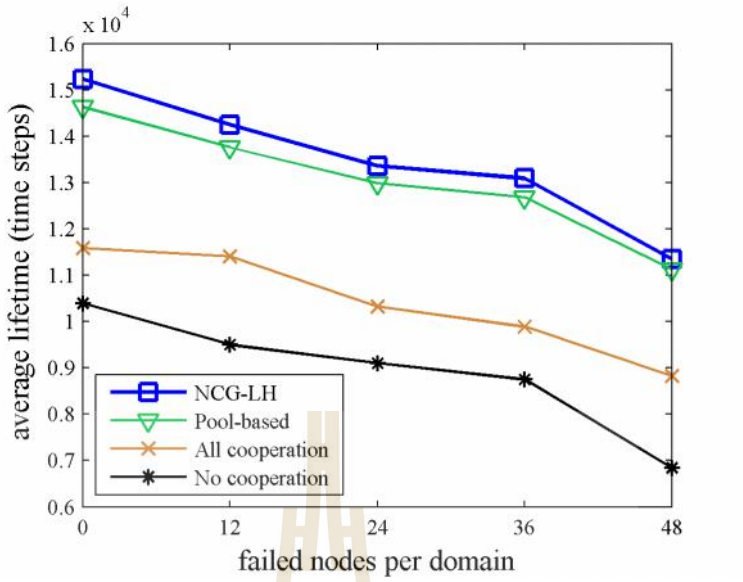
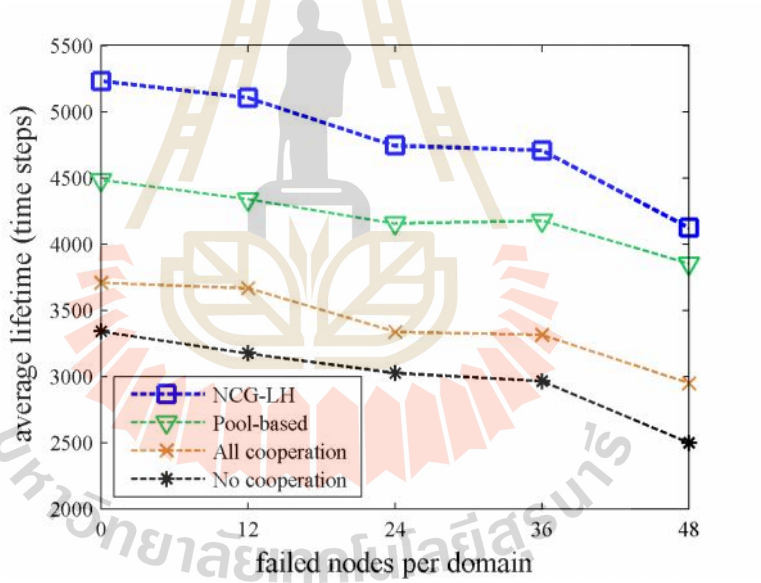


Figure 4.11 Average proportion of cooperation in various node failures under different path loss exponents



(a) free space



(b) PLE 4

Figure 4.12 Average network lifetime in various node failures and different path loss exponents

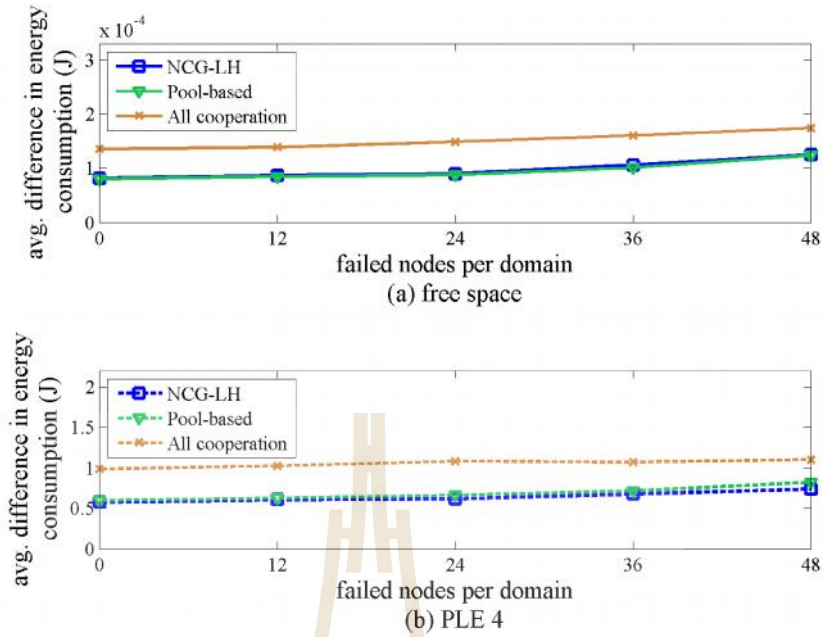


Figure 4.13 Average difference in energy consumption in various node failures and different path loss exponents

N_1 a N_2 a

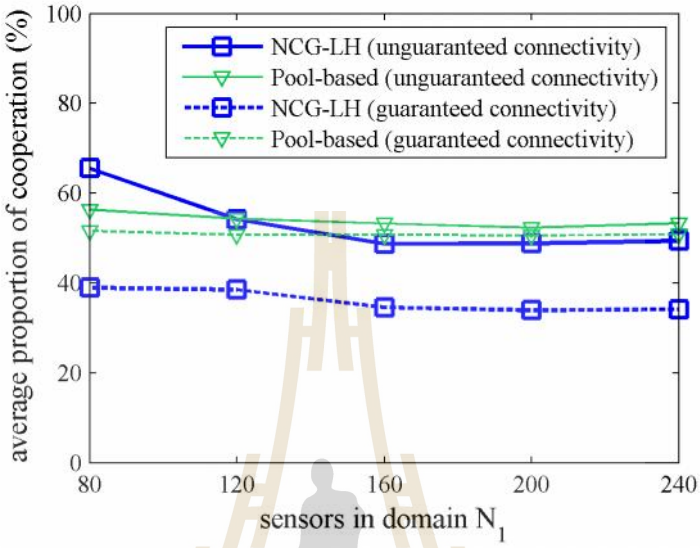


Figure 4.14 Average proportion of cooperation at different node density of network N_1

e av

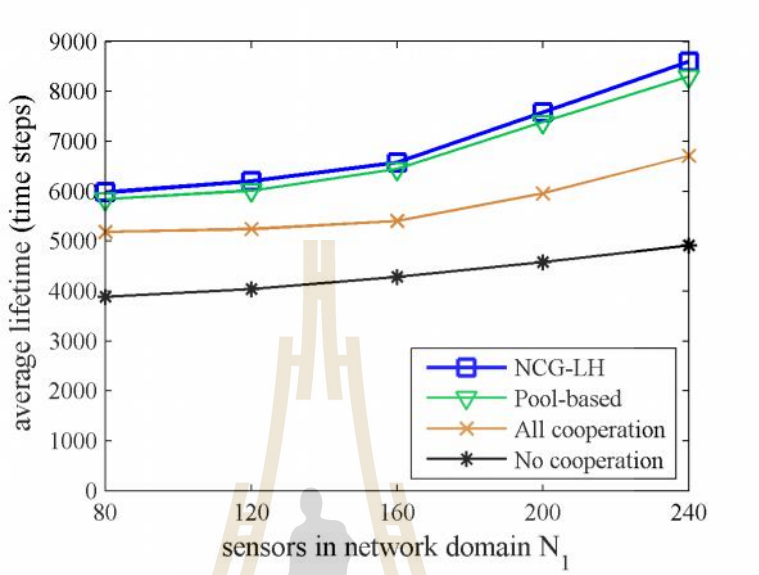
N_1. U

prop

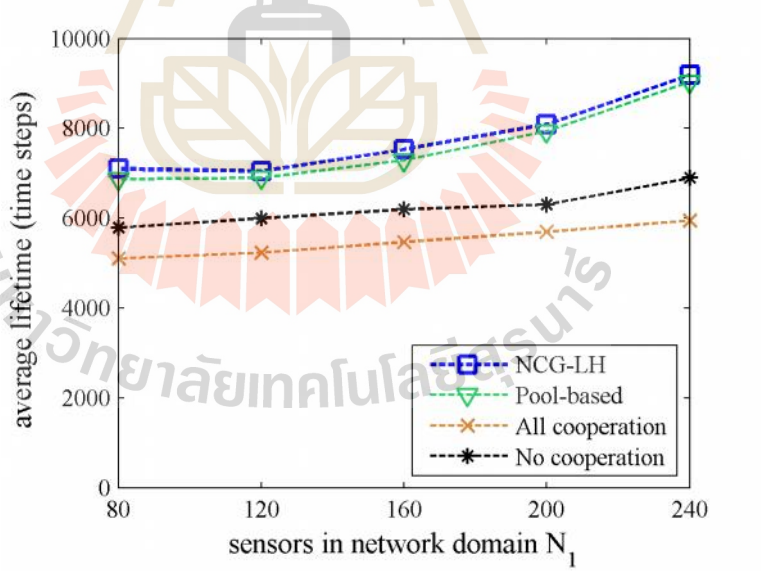
N_1 in

coope

N_1 in



(a) unguaranteed connectivity



(b) guaranteed connectivity

Figure 4.15 Average network lifetime at different node density of network N_1

ge di

N_1.

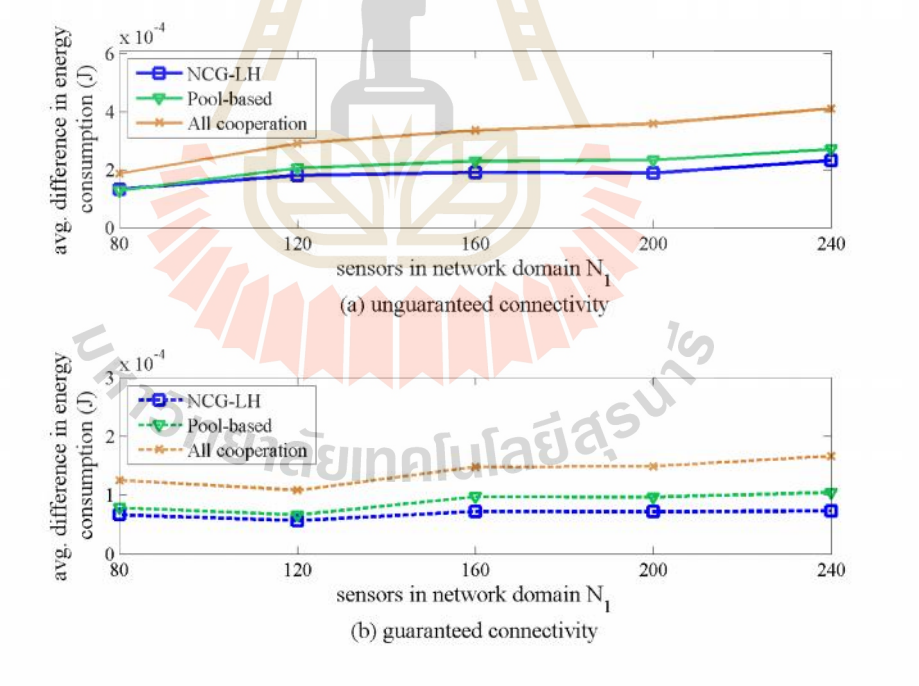


Figure 4.16 Average difference in energy consumption at different node density of network N_1

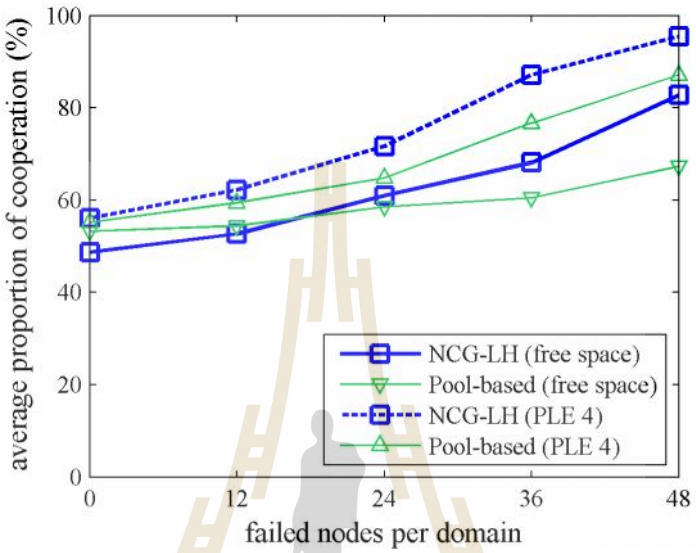
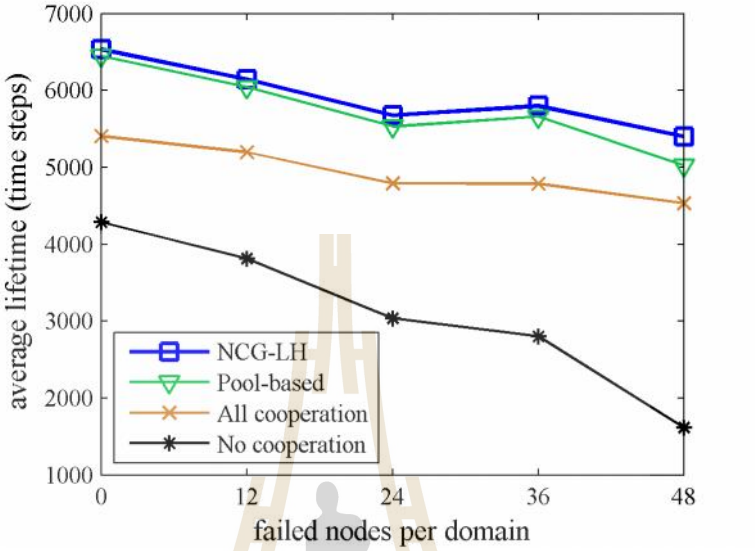
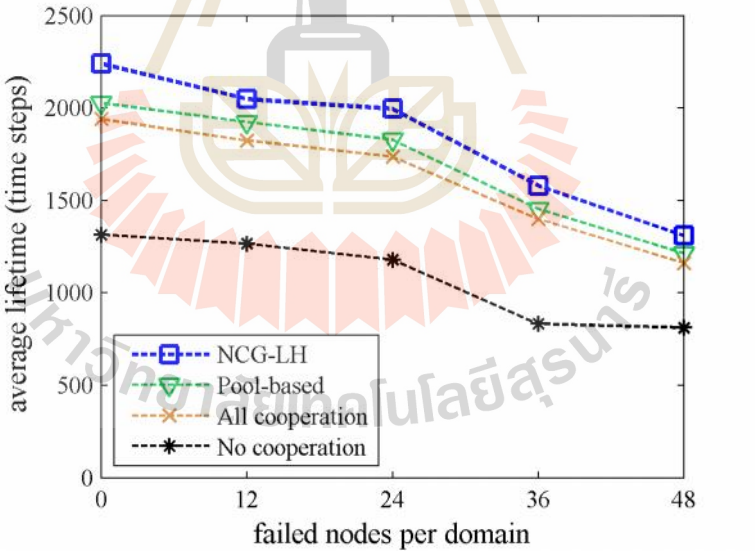


Figure 4.17 Average proportion of cooperation in various node failures under different path loss exponents



(a) free space



(b) PLE 4

Figure 4.18 Average network lifetime in various node failures under different path loss exponents

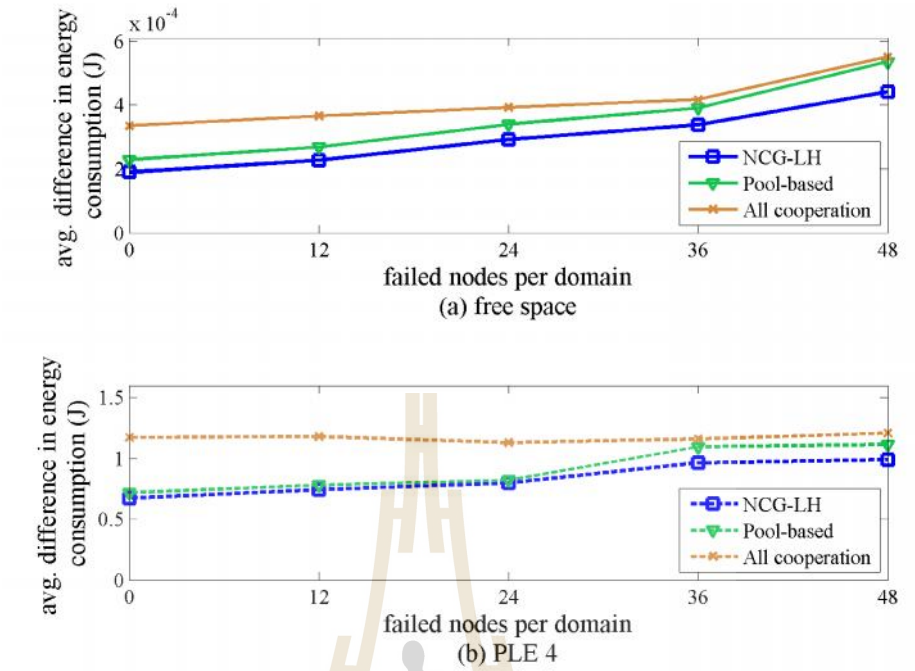


Figure 4.19 Average difference in energy consumption in various node failures under different path loss exponents

Secondly, investigation of fairness in terms of the difference in energy consumption between domains and comparison between a game theoretic approach (NCG-LH) and the non-game theoretic approach (Pool-based method). Finally, identification of parameters that effect cooperation between multiple co-located networks i.e., network density, node failure, PLE, network connectivity and sink positions.

The simulation results are divided into three scenarios. The study in scenario 1 is to investigate effect of cooperation in multi-domain WSNs with separate sink. The results show that when sink node in each WSN is separate, NCG-LH can promote more cooperation. Moreover, NCG-LH can obtain 4.3%-31.2% longer network lifetime than the other algorithms as network density, PLE and the number of failed node increases. Moreover, NCG-LH is comparable to Pool-based routing algorithm which promotes fair routing selection when compared to All cooperation algorithm.

In scenario 2, the difference in node density in each domain is studied (i.e. when number of sensors in domain N_1 are denser than domain N_2. NCG-LH can demote cooperation between domains due to the high availability of nodes and routes in domain N_1. This in turn, helps prolong network lifetime in domain N_2 which has less node density. The results show that NCG-LH obtains 3.3%-37.3% longer network lifetime than the others as network density, PLE and the number of failed node increases.

In scenario 3, the effect of sink position is studied. When the sink positions are moved further away from each other, NCG-LH promotes cooperation between networks compared to the original position and obtains 2.6%-39.1% longer network lifetime than the other algorithms as network density, PLE and number of failed node

increases. In addition, NCG-LH outperforms the other routing algorithms in terms of fair route selection by attaining the lowest average difference in energy consumption.

CHAPTER V

FAIR ROUTE SELECTION IN MULTI-DOMAIN WIRELESS SENSOR NETWORKS USING CONTINUOUS STATE NASH Q-LEARNING

5.1 Introduction

In multi-domain WSNs, cooperation among sensor nodes belonging different network authorities could potentially gain certain benefits. Such benefits include alternative routing paths and reduced energy consumption, which can prolong their network lifetime and enhance reliability of packet delivery. Most existing works focus on full cooperation in multi-domain WSNs (Bicakci et al., 2013; Jiang et al., 2013; Jelicic et al., 2014; Singhanat et al., 2015). All of these works showed that resource sharing and cooperation between sensor nodes in multiple domains, result in reduced energy consumption and increased network performance. However, Vaz et al., (2008) and Ze et al., (2012) showed that cooperation between two different networks that are deployed in the same region may not always be beneficial to both networks. This is because whether or not each sensor node will cooperate depends on the configuration of each network, network connectivity and how hostile the environment is. The previous chapters introduced the application of non-cooperative game theory to address this issue and proposed a routing algorithm named *Non-cooperative game algorithm based on Lemke Howson method* (NCG-LH) algorithm. The algorithm is

able to suitably determine packet forwarding strategy between multiple domains by using Nash equilibrium (NE) and Lemke Howson (LH) method. Note that this approach determines an action that maximizes only the *immediate payoff* in the current time step. An agent's choice of action results in a feedback (payoff or reward) and a change of state of system. A series of new actions and state changes thus given rise to a different accumulation of feedback (see Chapter 2). It is therefore interesting to investigate what happens if an agent can capture effects of actions beyond the next time step by maximizing the *expected future payoff* to get a suitable packet forwarding strategy in the current time step.

To address this issue, a model free tool called reinforcement learning (RL) has been introduced. In RL, agent can learn a behavior based on its reward (or payoff) value in the future time step to achieve the optimal strategy (Sutton and Barto, 1998). In the context of RL framework, an agent systematically learns correct behaviors online through trial-and-error interaction with other agents in order to achieve the action that maximizes its *expected future rewards*. There are several recent researches which employ RL to solve routing problems in WSNs (Kulkarni et al., 2011 and Al-Rawi et al., 2015). Each sensor node is assumed to be an agent. Therefore, WSNs with multiple independent decision-making agents can be considered as a multi-agent reinforcement learning (MARL) system. Recent researches applied a standard RL method called Q-learning to solve resource allocation problems in single domain WSN i.e. under a single network authority (Yang et al., 2013; Hu et al., 2010, Xu et al., 2015 and Debowski et al., 2016) formulated using MARL framework. Their results showed that their approach can maximize their network lifetime. On the contrary, limited research work have investigated in multi-domain WSNs with

networks controlled by multiple network authorities. Ref. (Rovcanin et al., 2014) considered scenarios of fully cooperative agents whereas (Singsanga et al., 2010) considered non-cooperative agents. However, both (Rovcanin et al., 2014) and (Singsanga et al., 2010) rely on a centralized operation, in which a single computational node (e.g. cluster head) receives and processes all sensor data, thus creating a large amount of overhead rendering it impractical for actual WSN applications. Hence, there is a need for decentralized or distributed algorithms that allow sensors to estimate their information locally to reduce the amount of overhead used.

Therefore, the objective of this chapter is to propose routing algorithms to deal with a non-cooperative multi-agent packet forwarding in multi-domain WSNs which is achieved by learning based on the expected future reward. It should be noted that learning based on future reward is considered in this chapter instead of immediate reward of NCG-LH from the previous chapter. The proposed algorithm is based on game theoretic reinforcement learning (GTRL) in order to select fair packet forwarding routes that can prolong network lifetime and enhance reliability for non-cooperative multi-domain WSNs in a distributed manner. Two routing algorithms are proposed in this chapter. The first algorithm is the *Discrete state Nash Q-learning* (*D-NashQ*), which is an extension of a centralized discrete state NashQ in (Singsanga et al., 2010) to support distributed multi-domain WSNs by using a payoff matrix derived in chapters 3 and 4 as reward function for the algorithm. The discrete state space is defined as the set of the actual battery levels of sensor nodes, which is divided into 3 levels. The other algorithm is the *Continuous state Nash Q-learning* (*C-NashQ*) that considers the state space as continuous state, which is suitable for the continuous state

of the remaining battery energy of sensor nodes. To the best of our knowledge, there is no existing work on applying GTRL for fair distributed packet forwarding problem in multi-domain WSNs. This chapter also evaluates the proposed algorithms and discusses their network performances. The results show that by using the proposed algorithms which provide fair route selection, all networks can send their packets more reliably and gain longer network lifetime.

The main contributions of this chapter are four-fold: 1) Derivation of feature function to represent the continuous state in continuous state Nash Q-learning; 2) Proposal of two distributed routing algorithms (D-NashQ and C-NashQ) and their application to the packet forwarding problem in multi-domain WSNs under separate sink scenario; 3) Comparison of Nash Q-learning performance in discrete state and continuous state; 4) Performance evaluation and comparison of C-NashQ and existing routing algorithms.

5.2 Related work

With the increasing use of WSNs technologies to a wide range of application scenarios, many researches tend to be more interested in resource allocation problem in multi-domain WSNs. This is due to cooperative resource sharing between multiple domain belonging different authorities which can reduce energy consumption and increase network performance.

Most existing researches consider resource allocation problem in a *cooperative* situation, meaning that, the network authorities have to agree on sharing or providing a common resource in order to increase the benefits of their networks. In ref. (Bicakci et al., 2013 and Bicakci et al., 2010), the potential benefits of

cooperation in multiple WSNs are investigated. Linear programming was employed to find energy efficient path in order to prolong their network lifetime. However, energy efficient route selection does not always guarantee a prolonged the network lifetime. Sensor nodes belonging to energy efficient paths tend to have higher traffic load and consume more energy than other nodes. As a result, such nodes tend to die earlier. In order to avoid heavily loaded situations. Nagata et al. (2012) proposed cooperation between multi-domain WSNs by balancing the communication load. Routes with the maximum value of bottleneck was selected. By doing this, network lifetime can be extended among multiple domains within the same geographic area. Kinoshita et al. (2016) proposed a fair cooperative routing method for heterogeneous overlapped WSNs called “Pool-based” selecting method. An energy pool was introduced to maintain the total amount of energy consumption by cooperative forwarding. Their simulation results showed that the proposed method was able to balance the energy consumption and prolong the network lifetime. Ref. in (Jelicic et al., 2014; Singhanat et al., 2015) showed benefits of node collaboration in multi-domain WSNs under practical implementation. The results showed that cooperation with co-located sensor devices in different networks can increase the network lifetime. In order to handle non-cooperative behaviors among sensor nodes in multi-domain WSNs, Wu and Shu (2005) applied the concepts from economics and game theory to propose a mechanism design (MD) approach. This approach is applied to a packet forwarding problem in multi-domain WSNs by using incentive mechanisms to motivate cooperation between sensor nodes. On the other hand, some researches employed non-cooperative game theory to the packet forwarding problem to describe such a situation that cooperation can exist in multi-domain WSNs without any incentive

mechanisms was proposed as both centralized algorithms. Ref. (Felegyhazi et al., 2005) showed that the Non-cooperative game algorithm is a suitable framework to determine an equilibrium strategy for their problem. However, one drawback of this approach is that obtaining a strategy needs significant amount of computational time to compute the utility for all possible actions of sensor nodes. A two-agent relaying game was analyzed in a centralized non-cooperative game framework under separate sink scenario was proposed in (Yang and Brown, 2007). However, their experiment investigated a small network with two sensors and two separate sinks.

In this chapter, we introduce the application of multi-agent reinforcement learning (MARL), which is another technique to address the issue of resource allocation problem in WSNs. MARL is suitable for distributed routing problems. A standard RL method called, Q-learning has been proposed to determine best routing strategies when critical network conditions are allowed to vary dynamically. In (Yang et al., 2013), a MARL-routing approach was proposed to handle sink mobility and enable direct interactions between WSN and vehicles. Reward functions including time delay, network lifetime and reliability was designed for learning. Simulation results showed that their proposed approach achieved better time delay, energy distribution and delivery rate than comparing routing approaches. Refs. (Hu et al., 2010, Xu et al., 2015 and Debowski et al., 2016) presented a load-balancing multi-path routing approach. A MARL technique was employed to learn and find out the best path to forward packet which considers the number of hops, residual energy and energy consumption of sensor nodes. Results showed that their approaches can balance the workload among sensor nodes and prolong the network lifetime. However, these solutions were directly applied in single-domain WSNs. There are

only a few researches focus on MARL technique in multi-domain WSNs. Ref. (Rovcanin et al., 2014) applied Q-learning to solve routing problem for cognitive networks such networks were co-located heterogeneous WSNs which were fully cooperative operating in a centralized manner. MARL under centralized manner was also proposed in (Singsanga et al., 2010), by extending Q-routing to cater a non-cooperative multi-agent in a packet forwarding problem. The authors applied an existing algorithm called Nash Q-learning (NashQ) previously proposed in (Hu and Wellman, 2003) to attain the best mutual policy for all agents in a packet forwarding game framework. Each agent attempts to learn its Nash equilibrium (NE) online. Their results suggest that NashQ can learn and determine a suitable packet forwarding policy in varying network conditions. Moreover, both (Rovcanin et al., 2014) and (Singsanga et al., 2010) rely on a centralized operation, which was impractical for storage and computing ability of sensor nodes. Therefore, to the best of our knowledge none of the existing GTRL researches take into consideration of fair routing selection in multi-domain WSNs in a distributed manner. Because in a multi-domain environment, lifetime improvement with cooperation may not be fair to all domains. It is possible that some WSNs can prolong their network lifetime but for other WSNs, their network lifetime may be reduced. Therefore, for fair cooperative routing, it is necessary to take into consideration the energy that sensors in each domain consume in packet forwarding.

This chapter therefore proposes a fair distributed packet forwarding algorithm in multi-domain WSNs based on GTRL. In particular, *D-NashQ* and *C-NashQ* algorithms are proposed in this chapter in order to learn a fair packet forwarding policy based on discrete and continuous battery states. The two algorithms are based

on the game theoretic reinforcement learning (GTRL) technique to support non-cooperative behaviors of sensor nodes belonging to different networks. An introduction to GTRL technique is described in the next section.

5.3 Game theoretic reinforcement learning

5.3.1 Reinforcement learning

Reinforcement learning (RL) (Sutton and Barto, 1998) is a machine learning scheme to provide a framework in which an agent can learn optimal control policy based on the agents' past experiences and reward. RL relies on the assumption that the dynamics of the system follows a *Markov Decision Process* (MDP) (See Chapter 2). A MDP models an agent acting in an environment with a tuple (*S, A, P, R*), where *S* is the set of states, *A* is the set of actions that the agent could take in a particular state, *P* is the state transition probability matrix and *R* is a function of reward expected from the environment after taking the action $a \in A$ at state $s \in S$ and transiting to next state $s' \in S$. In MDP, the objective is to find a policy $\pi : S \rightarrow A$ which is a mapping of the state set to the action set through interacting with environment to maximize objective function.

5.3.2 Q-learning

A common RL technique called Q-learning is employed to solve for an optimal policy in MDPs in single-agent systems. It is a model-free online learning method that has been applied widely because of its simplicity. It can effectively make an agent to learn optimal policy through trial and error and can directly converge to the optimal *action-value function* (Q-value) through online learning. The key point of Q-learning is updating Q-value through iteration. An agent takes action *a* at state *s*,

receives reward *r*, updates the local state with input from the environment, and repeats the process to learn its own optimal policy. Q-learning provides a simple procedure in which the agent starts with an arbitrary initial Q-value at time step *t=0*. The updating process at time step t+1 is defined as

$$Q_{t+1}(s,a) = (1-\sim)Q_t(s,a) + \sim[r_t + \max_{a'} Q_t(s',a')], \quad (5.1)$$

where $\sim \in [0,1)$ is the learning rate parameter, $\in [0,1)$ is the discount factor, r_t is reward at time t and s' is the next state that results from taking action *a* in state *s*. Several researches i.e. (Yang et al., 2013; Hu et al., 2010, Xu et al., 2015 and Debowski et al., 2016) employ Q-learning to solve routing problems in single-domain WSNs. Each sensor node is modeled as an agent and then the entire wireless sensor network can be modeled as a multi-agent reinforcement learning (MARL) system. In order to select the optimal path, each agent selects an optimal neighbor node as its next hop to forward its data packet to its sink. Their results show that this method can improve network performances in their system by taking advantage of cooperative behavior of sensor nodes. However, Q-learning cannot be directly applied to multi-domain WSNs as cooperative behavior between sensor nodes belonging to different domains may not always be available. This is due to selfish behavior of sensor nodes in different networks domain to conserve energy for their own network. Therefore, the optimal policy for a WSN does not only depend on one domain, but also other domains located in the same region.

5.3.3 Nash Q-learning

Hu and Wellman (2003) proposed algorithm called Nash Q-learning (NashQ), an extension of the original Q-learning to a non-cooperative multi-agent

system. In NashQ algorithm, each agent can rationally decide its own action whether it will cooperate with other agents or not by considering both its own and other agents' information as well. Instead of finding an optimal policy to maximize one single agent's reward like the original Q-learning, NashQ looks for joint actions that yield the best reward for all agents. The agents attempt to learn their best mutual policy, which is defined by the Q-values received from Nash equilibrium (NE). NE is not only used to decide the agent's own action policy, but also predict the other agent's action policy, given by $®^1(s'),...,®^v(s')$ where $®^i(s')$ is agent i's distribution over its set of actions at state s' and v is the number of agents. NE can be found in a pure-strategy equilibrium, where an agent is able to find the highest mutual utility for all agents. But in general, not all games have pure-strategy NE. The agents then have to then decide whether to select their policies randomly according to some calculated probability to achieve the best response. Such NE behavior is called mixed-strategy Nash equilibrium. The Lemke-Howson method (LH) is the best known method to solve for mixed-strategy NE for two agents (Shoham and Brown, 2009). The advantage of LH method is that it is guaranteed to find at least one NE point.

In this chapter, we thus employ NashQ into packet forwarding problem in a non-cooperative multi-domain WSNs in order to find the best mutual policy which provides the best benefits for all agents in the system. The source node, which is randomly selected from sensor nodes in the WSN, is modeled as an agent. The entire WSN can thus be modeled as a multi-agent system. In order to select the optimal packet forwarding path, each agent selects a fair routing obtained from NE.

Two routing algorithms are proposed in this chapter. The first algorithm is the *discrete state Nash Q-learning* (D-NashQ), an extension of centralized discrete

state NashQ in (Singsanga et al., 2010) to support a distributed multi-domain WSNs. A payoff matrix derived in chapter 3 is used as a reward function. In D-NashQ, Q-value functions are estimated in tabular forms for each state or state-action pair. However, many real-world applications have to deal with MDPs with continuous state spaces. So Q-learning in discrete state may not be feasible. In such cases, another algorithm called the *continuous state Nash Q-learning* (C-NashQ) is proposed. C-NashQ learns policy by using a proposed feature function that is suitable for continuous state.

This section briefly introduces the application of RL, original Q-learning and NashQ to address the issue of non-cooperative resource allocation problem in multi-domain WSNs. The routing problem for multi-domain WSNs is modeled based on NashQ present in the next section.

5.4 Routing model based on NashQ approach

The objective of using NashQ algorithm in this chapter is to select an online fair packet forwarding policy in multi-domain WSNs. The proposed algorithm was then designed by considering communication cost in multiple route paths in order to provide maximum savings in energy and network lifetime. Furthermore, the residual energy of sensor node must be taken into account in order to balance the network load to achieve fairness. The NashQ algorithm can thus efficiently determine packet forwarding policy to obtain fair energy consumption for all network domains, prolong the network lifetime and enhance reliability of packet forwarding. *D-NashQ* and *C-*

NashQ routing algorithms are modeled based on GTRL technique which are presented below.

5.4.1 Network model

Consider two different WSNs, N_i, $i=1,2$, deployed in a multi-domain WSN. Each WSN domain consists of v sensor nodes, $N_i = \{n_i^1, n_i^2, \ldots, n_i^v\}$, and one sink. Our model divides the time into discrete time units called time steps. In each time step, a source node is randomly selected from sensor nodes in each domain to generate a data packet and send it to its sink. The source node acts as an agent which decides a route to send the data packet by using the proposed routing algorithms. This chapter assumes the following characteristic of each sensor node in the multi-domain WSNs.

- Two sensor nodes are able to communicate with each other if they are within transmission range.
- Each sensor node must be aware of its location, neighbor location, its sink location and also neighbors sink location using an on-board GPS receiver.
- There is a pre-established routing mechanism using AODV routing protocol to determine two routes: 1) a route that contains nodes from the same domain as source node and 2) another route containing multi-domain nodes.
- The source node is able to calculate the cost of a transmission, which is the end-to-end distance from source node to sink.
- The total energy consumption of each sensor node are dissipated only for data transmission and reception.

The energy consumption required for packet forwarding is computed from the radio model in (Naruephiphat and Usaha, 2008). The radio model for the reception cost of each sensor node is given by, $E_{RX}(b) = E_{elec} \times b$, where $E_{elec} = 50\ nJ / bits$ is the cost in the radio electronics and we assume that b is the size of the measurement packet transmitted in bytes. The transmission cost is for each sensor node given by, $E_{TX}(b,d) = (E_{elec} \times b) + (£_{amp} \times b \times d^{±})$ where σ is the path loss exponent and $£_{amp} = 10\text{pJ} / \text{bit} / m^2$ is the energy consumed at the output transmitter antenna for transmission range of one meter.

In a pre-established routing process, AODV routing protocol, which is used in IEEE standard 802.15.4 ZigBee protocol stack (ZigBee Alliance, 2015), is employed to establish available route paths. In the route discovery process, source node broadcasts Route Request (RREQ) packets to its neighbors in the same domain and also neighbors in difference network domain. The source node then establishes two different routes with two different routing tables, one for routing within the source node's own network and the other for coordinating paths with the other network domain (as shown in Figure 5.2).

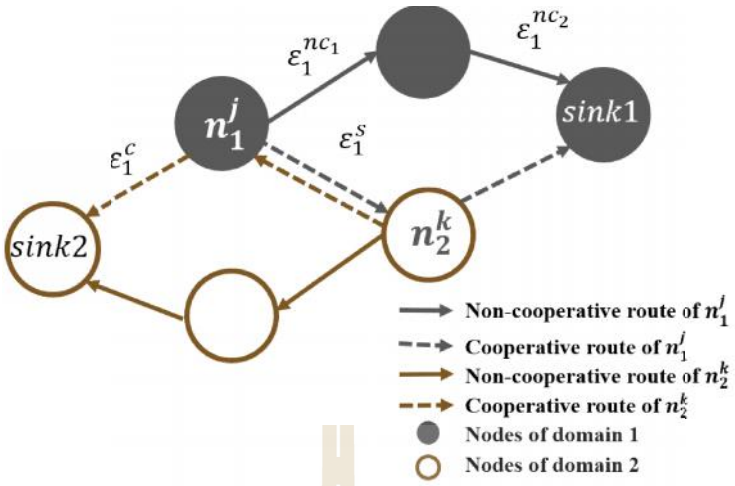


Figure 5.2 System Model

From the figure, sensor node n_1^j where $j=1,2,\ldots,v$ from network domain N_1 is randomly chosen to be a source node taking a role as an agent in the game at time step t. When n_1^j has a packet to send to its sink1, n_1^j must decide whether to use the non-cooperative route in its own domain or the cooperative route that consists of nodes from the other domain. To make a decision, n_1^j calculates the following energy, including: 1) the end-to-end energy cost along non-cooperative route, $£_1^{nc} = £_1^{nc_1} + £_1^{nc_2}$; 2) the energy required at n_1^j to forward domain N_1's packets through domain N_2's node (i.e. forward its packet to sensor node n_2^k, where $k=1,2,..v$) to sink1, $£_1^s$; and 3) the end-to-end energy used by nodes in domain 1 required to help domain N_2 forward domain N_2's packets to sink2, $£_1^c$. These energy values are used to estimate a payoff value that an agent receives in order to decide which packet forwarding path to choose. The payoff value is used as a reward function described in section 5.4.2. The optimal packet forwarding path will be

chosen by the source node depending on strategy decision through *D-NashQ* and *C-NashQ* algorithm, described respectively in section 5.4.3 and 5.4.4.

5.4.2 Action formulation and reward function

For the sake of simplicity, consider a WSN of 2 domains. Let *A* be an action space defined as a set of strategies, which include all the possible joint strategies or actions available in the game. Let the action space for domain N_i be is defined by $A_i = \{D, F\},\ i = 1, 2$, where the shorthand notations refer to the following:

D: The agent *does not forward* its packet to the other network (i.e. agent chooses the non-cooperative route) and *drops* all packets from other network if asked for help to forward the packets.

F: The agent *forwards* its packet to the other network (i.e. agent chooses the cooperative route) and in turn forwards all packets if the other network asked for help to forward the packets.

Therefore, the set of joint actions for agent in both domains is *{DD, DF, FD, FF}*. The reward function is the feedback from taking a joint action of agent. The reward function is significant since the objective of learning is to achieve an optimal policy with the maximum reward. For WSNs, it is a better approach to enable nodes to not only reduce energy consumption whenever possible, but also transport the data reliably to a sink. Therefore, reward function can be take energy consumption and link quality into consideration. A reward function according to the payoff matrix (derived in Chapter 3) is proposed in Table 5.1. In this table, the agent from domain N_1 is the row agent and the agent from domain N_2 is the column agent. Thus from Figure 5.2, n_1^j is the row agent and n_2^k is the column agent. Each agent has two actions

i.e. to forward (*F*) or do not forward (*D*) the packet to the other agent. Each cell of the matrix contains a pair of values represent the reward of agent n_1^j and n_2^k. The first value is the reward of agent n_1^j and the second value is the reward of agent n_2^k. For example, at time step *t*, if agent n_1^j and n_2^k take action *D* and *D* respectively, agent n_1^j receive reward $r_1^t = «_1$, whereas agent n_2^k receive reward $r_2^t = «_2$.

Table 5.1 Reward function of interaction between sensor nodes in different domains

	D	**F**
D		$0, -\frac{F}{\eta_2}$
F	$-\frac{}{\eta_1}, 0$	$\frac{}{\mu_1}(\frac{}{\gamma_1} - \frac{}{\eta_1}), \frac{}{\mu_2}(\frac{}{\gamma_2} - \frac{}{\eta_2})$

The parameter μ_i is the packet received rate (PRR) (Ahmedand and Fisal, 2008), which is approximated as the probability of successfully receiving a packet from source node to sink, and $i = 1,2$. The higher PRR is, the higher the link quality is. The PRR can be calculated from the bit error rate (BER) follows:

$$PRR = (1 - P_b)^b, \tag{5.2}$$

where P_b is the bit error probability for OQPSK (Offset Quadrature Phase Shift Keying) modulation used in IEEE standard 802.15.4 ZigBee protocol stack at frequency 2.4 GHz. The other parameters in Table 5.1 refer to Figure 5.2. The quantity $¥_i = £_i^{nc} - £_i^{s}$ denotes the energy reduction obtained from changing from the non-cooperative route to the cooperative route. The quantity $¦_i = £_i^{c}$ is the

cooperative energy required for cooperation. Finally, the quantity ¥$_i$ − ¦$_i$ is the net energy gain if the source node chooses the cooperate route. If ¥$_i$ − ¦$_i$ is a positive value, it means that the cooperative route consumes less energy than the non-cooperative route. Otherwise, the cooperative path consumes more energy.

5.4.3 D-NashQ approach

In the context of online learning, each agent decides its state *s*, gets an immediate reward *r* and update the Q-value. In general, the state space usually is defined as a discrete state. The discrete state, Q-value and the updating of Q-value for D-NashQ are defined as follows.

5.4.3.1 Discrete state definition

In D-NashQ, the state space is defined as the set of the discrete battery energy of the sensor nodes. Since the battery energy is continuous, we divide the range of battery energy of each agent into 3 states given by $S = \{0, 1, 2\}$. Initially, the state of each agent is state, “2” meaning full battery level. The game is repeated until the any agent reaches state “0”, signifying battery depletion of a sensor node in its domain and the game then ends.

5.4.3.2 Q-value and Q-updating

Through learning, an agent can updates its Q-value which represents to the reward of each action in a particular state. The optimal forwarding path then can be selected by choosing from the best mutual Q-value as follows. At the beginning, the Q-value functions are initialized to $Q_i^0(s_i, a_1, a_2) = 0$, for all $s_i \in S_i$, $a_i \in A_i$, $i = 1, 2$. Let the learning agent be indexed by *i*=1. Upon a packet

transmission, at time step *t*, agent n_1^j (assumed as source node in domain N_1 at that time step) observes the current discrete state, takes its action by selecting $a_1 = D$ or F, and observes its own reward. It then observes the action, reward at the other agent and observes the next state $s_1' \in S_1$, of both agents. Agent n_1^j then calculates a NE policy $\mathbb{R}_1(s_1'), \mathbb{R}_2(s_2')$, where $\mathbb{R}_i(s_i')$ is distribution of agent in domain N_i over its set of actions at state s', for the stage game $(Q_1^t(s_1, a_1, a_2), Q_2^t(s_2, a_1, a_2))$ and updates its Q-values as follows

$$Q_1^{t+1}(s_1, a_1, a_2) = (1 - \sim) Q_1^t(s_1, a_1, a_2) + \sim [r_1^t + \quad NashQ_1^t(s_1', a_1', a_2')], \tag{5.4}$$

where $NashQ_1^t(s_1', a_1', a_2') = \mathbb{R}_1(s_1') \cdot Q_1^t(s_1', a_1', a_2') \cdot \mathbb{R}_2(s_2')$. (5.5)

$NashQ_1^t(s_1', a_1', a_2')$ is agent n_1^j's Q-values in state s' for the selected NE. Note that $\mathbb{R}_1(s_1') \cdot Q_1^t(s_1', a_1', a_2') \cdot \mathbb{R}_2(s_2')$ is a scalar. For any stage game, at least one NE exists in either pure or mixed strategies. In pure strategy NE, an agent can choose with certainty join action with highest Q-values for itself and the other agent. The method for selecting mixed strategy NE is the Lemke-Howson method (see appendix A).

In order to calculate the NE strategy, agent n_1^j must observe the other agent's information (i.e. agent n_2^k, which is the sensor node in a different network domain belonging to cooperative route as seen in Figure 5.2) that are immediate

reward and previous actions and updates its conjecture on the other agent's Q-function, by maintaining its own update on the other agent's Q-function

$$Q_2^{t+1}(s_2,a_1,a_2) = (1-\sim)Q_2^t(s_2,a_1,a_2) + \sim[r_2^t + \quad NashQ_2^t(s_2',a_1',a_2')], \qquad (5.6)$$

where $NashQ_2^t(s_2',a_1',a_2') = \mathbb{R}_1(s_1') \cdot Q_2^t(s_2',a_1',a_2') \cdot \mathbb{R}_2(s_2').$ (5.7)

In D-NashQ, we set the learning rate parameter, $\sim = 0.01$ and the discount factor $= 0.01$ (see Appendix B).

It can be seen that the agent in NCG-LH algorithm in chapters 3 and 4 only seek a strategy that ontained from NE, $NashU_i^t = \mathbb{R}_1 \cdot U_i^t \cdot \mathbb{R}_2$, where $NashU_i^t$ is agent's payoff value for the selected NE point and U_i^t is a payoff matrix of agent *i* at time step *t* (Table 3.1). On the other hand, the agent in NashQ algorithm learns to get a strategy based on Q-values in state s' for the selected NE, $NashQ_i^t(s_i',a_1',a_2')$. Moreover, $NashQ_i^t(s_i',a_1',a_2')$ value is used in improving its own Q-table by updating following Eq. (5.4) in order to determine a strategy in the next time step.

5.4.3.3 Mutual policy

The Q-learning involves finding a balance between exploration strategy and exploitation strategy. Each agent uses the $£$-*greedy* method to select its action. In this method, each agent selects the NE action with probability $1-£^t(s)$ (so called exploitation) and selects an action randomly with probability $£^t(s)$ for other Non-NE action (so called exploration). $£$-greedy probability, $£^t(s)$ is defined as

$$£^{t}(s) = \frac{1}{1 + 0.1K^{t}(s)}, \tag{5.8}$$

where $K^{t}(s)$ is number of visits to state *s* at time *t* of agent *i*. The pseudo code of D-NashQ is shown in Figure 5.3.

BEGIN

for topology 1:100

Initialize energy and initial state s_0 for each node to full battery level

Let the learning agent be indexed by *i*.

Let $Q_i^t(s_i, a_1, a_2) = 0$ for *i*=1,2

Let t=0

do

Random source node to create data packet

Establish two routing tables using AODV routing protocol (one table for paths in own network and another one for paths in cooperative networks)

Take action $a_1,\ a_2$, receive reward $r_1,\ r_2$ and next state $s_1',\ s_2'$

Update $Q_i^{t+1}(s_i, a_1, a_2)$ for *i*=1,2

$$Q_i^{t+1}(s_i, a_1, a_2) = (1 - \sim)Q_i^t(s_i, a_1, a_2) + \sim[r_i^t + \quad NashQ_i^t(s_i', a_1', a_2')]$$

Let t=t+1

while (at least one node run out of battery)

endfor

END

Figure 5.3 Pseudo code of D-NashQ algorithm

5.4.4 C-NashQ approach

Most RL routing techniques are often modeled as Markov Decision Processes (MDPs) with discrete state and action spaces to simplify the use of RL algorithms to find solutions. However, real world problems may have continuous state spaces. This chapter defines the state space as the set of actual battery energy of the

sensor nodes. Since the state value is continuous, quantizing continuous values to discrete values may obtain suboptimal policies during the learning process. To address this constraint, we extend the original NashQ algorithm to continuous state spaces context and proposed *continuous state Nash Q-learning (C-NashQ).* C-NashQ approach can learn near-optimal packet forwarding policy that maps a continuous state space to discrete action space. The objective remains the same as D-NashQ, which is to prolong network lifetime and enhance reliability of packet forwarding and obtain fair energy consumption for all network domains in multi-domain WSNs with distributed manner.

5.4.4.1 Continuous state definition

The continuous state is defined by a feature function $¤_i : S_i \times A_1 \times A_2$ where *i=1,2,* which maps a state-action pair to a particular function. Let $\mathbf{\Phi} = [¤_i(s_i, a_1, a_2)]$ be a feature function matrix. $\mathbf{\Phi}$ is a matrix of $|A_1| \times |A_2|$ dimension. Each element of the feature matrix $\mathbf{\Phi}$, $¤_i(s_i, a_1, a_2)$, is called a *feature*. Let $¤_i(s_i, a_1, a_2)$ be the feature value of agent in domain N_i for state-action pair (s_i, a_1, a_2) The characteristic of a good feature is that it should be able to represent states that continuously respond to changing actions. Thus, feature proposed in this chapter is a function which models the remaining energy after taking an action in the packet forwarding process, and is expressed by

$$¤_i^t(s_i, a_1, a_2) = ¨(route) \times \left(\frac{E_{remain}(s_i, a_1, a_2) - E_{total}(s_i, a_1, a_2)}{E_{initial}} \right), \quad (5.9)$$

where $\varphi(\cdot)$ is an indicator function which is defined by

$$¨(route) = \begin{cases} 1 \ , if \ \ a \ route \ associated \ to \ a_1, a_2 \ \ is \ available \\ 0 \ , otherwise. \end{cases} \quad (5.10)$$

The parameter E_{remain} is the remaining battery energy for all nodes in the route, E_{total} is the total energy consumption for packet forwarding in the route and $E_{initial}$ is the initial battery energy of sensor nodes in the route. The quantity $E_{remain}(s_i, a_1, a_2) - E_{total}(s_i, a_1, a_2)$ is the remaining energy after taking an action in packet forwarding process which is normalized by $E_{initial}$. The quantity $\varphi(\cdot)$ indicates the route presence for the agent. If the agent has an available route associated to action a_1, a_2 to send its packet, then ¨ $(route) = 1$, meaning that the remaining energy after taking an action can be determined only when such route exists. Otherwise, ¨ $(route) = 0$.

5.4.4.2 Q-value and Q-updating

In C-NashQ, Q-values can be approximated as a feature function (Geramifard et al., 2013):

$$Q_i(s_i, a_1, a_2) = ¤_i(s_i, a_1, a_2)^{-}{}_i, \tag{5.11}$$

where $\theta_i \in \mathbb{R}$ is a weight value of agent in domain N_i to be adjusted (see Figure 5.4) in order to achieve a NE point in the action value functions. The action value function can also be presented in a matrix form given by:

$$\mathbf{Q} = \mathbf{\Phi}^{-}, \tag{5.12}$$

where $\mathbf{\Phi}$ is the matrix of feature of dimension $|A_1| \times |A_2|$.

At the beginning, the weights of each agent *i* are initialized to $^{-}_{i}{}^{0} = 0,$ for all $s_i \in S_i,\ a_i \in A_i,\ i = 1,\ 2.$ Let the learning agent be indexed by *i*=1. At time step *t*, agent n_1^j (assumed as the source node in domain N_1 at that time step) observes the current continuous state, takes its action by choosing a neighboring node to forward a packet to and observes its own reward. It then observes the action, reward at the other agent and observes the next state of both agents. Agent n_1^j then calculates the Nash equilibrium strategy and updates its Q-values as follows

$$Q_1^{t+1}(s_1, a_1, a_2) = r_1^t + \quad NashQ_1^t(s_1', a_1', a_2'), \tag{5.13}$$

where $NashQ_1^t(s_1', a_1', a_2') = ®_1(s_1') \cdot ¤_1(s_1', a_1', a_2')^{-}_{1} \cdot ®_2(s_2').$ (5.14)

It can be seen that, when Q-values in C-NashQ is estimated be a feature function (Eq. 5.11). In particular, $NashQ_1^t(s_1', a_1', a_2')$ in D-NashQ (Eq.5.5) is changed to Eq. (5.14). Then, agent n_1^j needs to calculate ¢ , the temporal difference (TD) error, which is the difference of Q-value in the previous time step and the current time step

$$¢_1^t = Q_1^{t+1}(s_1, a_1,\ a_2) - Q_1^t(s_1, a_1,\ a_2). \tag{5.15}$$

The feature function $¤_i(s_i, a_1, a_2),$ calculated from $^{-}$ which is the weight specifying the contribution of each feature across all state-action pairs.

$$^{-}_{1}{}^{t+1} = ^{-}_{1}{}^{t} + \sim ¢_1^t ¤_1^t(s_1, a_2,\ a_2). \tag{5.16}$$

Agent n_1^j then observes the other agent's immediate reward and previous actions and updates its conjecture on the other agent's Q-function, by maintaining its own update on the other agent's Q-function

$$Q_2^{t+1}(s_2, a_1, a_2) = r_2^t + \quad NashQ_2^t(s_2', a_1', a_2'). \tag{5.17}$$

The temporal difference (TD) error, ¢ , can be determined by

$$¢_2^t = Q_2^{t+1}(s_2, a_1, a_2) - Q_2^t(s_2, a_1, a_2). \tag{5.18}$$

The parametric weight, ¯ , can be updated according to

$$¯_2^{t+1} = ¯_2^t + \sim ¢_2^t ¤_2^t(s_2, a_2, a_2). \tag{5.19}$$

In C-NashQ, we set the learning rate parameter, $\sim = 0.1$ and the discount factor $= 0.25$.

5.4.4.3 Best mutual policy

Both D-NashQ, C-NashQ also use the ε-greedy method to select actions for each agent. However, when discrete state is not considerated in this model, £$^t(s)$ (from Eq. 5.8) then changed to £$^t(a)$. Each agent selects the NE action with probability 1-£$^t(a)$ for exploitation, and selects an action randomly with probability £$^t(a)$ for exploration policy. £-greedy probability, £$^t(a)$ is defined as

$$£^{t}(a) = \frac{1}{1 + 0.1K^{t}(a)}, \tag{5.20}$$

where $K^{t}(a)$ is number of time action a is selected at time t of agent i. The pseudo code of C-NashQ is shown in Figure 5.4.

5.4.5 Compared algorithms

In order to evaluate the performance of the proposed routing algorithm, we compared it with 4 routing algorithms which include 1) NCG-LH algorithm which is proposed in Chapter 3 and 4. NCG-LH is an algorithm that determines packet forwarding policy by using (non-learning) non-cooperative game theory; 2) Pool-based routing algorithm (Pool-based) proposed in Kinoshita et al. (2016). This is a (non-learning) load balancing routing algorithm for multi-domain WSNs; 3) a classical (non-learning) AODV routing schemes which uses AODV to discover a route consisting of nodes within the same domain (No cooperation); and 4) a classical (non-learning) AODV routing schemes which discovers a route that consist of nodes from the other domain (All cooperation).

BEGIN

for topology 1:100

Initialize energy for each node to full battery level

Let the agent be indexed by *i*.

Let t=0

$\pi_i^0 \leftarrow$ Initialize arbitrarity

do

Random source node as the agent to create data packet

The agent establish two routing tables using AODV routing protocol

(one table for paths in own network and another one for paths in cooperative networks)

Take action a_1, a_2 , receive reward r_1, r_2 and next state $s_i' = \pi_i^{t+1}(s_i', a_1, a_2)$ for i=1,2

Update Q_i , α_i and π_i

$Q_i^{t+1}(s_i, a_1, a_2) = r_i^t + \quad NashQ_i^t(s_i', a_1', a_2')$

where $NashQ_i^t(s_i', a_1', a_2') = \phi_1(s_1') \cdot \pi_i(s_i', a_1', a_2')_{\pi_i} \cdot \phi_2(s_2')$

$\alpha_i^t = Q_i^{t+1}(s_i, a_1, a_2) - Q_i^t(s_i, a_1, a_2)$

$\pi_i^{t+1} = \pi_i^t + \sim \alpha_i^t \pi_i^t(s_i, a_2, a_2)$

Let t=t+1

while (at least one node run out of battery)

endfor

END

Figure 5.4 Pseudo code of C-NashQ algorithm

5.5 Simulation results

In this section, we evaluate the performance of two proposed GTRL routing algorithms, D-NashQ and C-NashQ, and investigate the cooperative conditions of the packet forwarding strategies in multi-domain WSNs. We use visual C++ to simulate the proposed routing algorithms. We consider two WSNs existing in the same area and the simulation environment is set to be a square area with 2500 m^2. Each WSN domain consists of one sink and 20-100 sensor nodes are deployed randomly. In each

time step, each WSN randomly chooses a source node to send data packets to its sink. The source node acts as agent of the packet forwarding game. The objective determine fair routing policy in order to prolong network lifetime and enhance reliability in distributed multi-domain WSNs by using the proposed algorithms.

Simulations are then carried out under varying number of nodes, number of failed nodes and path loss exponent. We compare the proposed algorithm with 4 existing algorithms considering 4 metrics including:

- *Proportion of cooperation:* the ratio of the number of cooperative routes to the total number of routes discovered.
- *Packet delivery ratio (PDR):* the ratio of the number of data packets received over the number of data packets sent out.
- *Network lifetime:* The lifetime of each network. Since each time step, a packet is transmitted, this chapter thus measures the network lifetime in terms of the total number time steps that data packet is transmitted at the sink node until the first node dies.
- *Fairness:* the difference in average energy consumed along a forwarding path between network domain N_1 and N_2.

The simulation parameters are shown in Table 5.2. Simulation results were carried out over 100 randomly topologies. The experimental results are shown in this section obtained from average results from both domains.

Table 5.2: Parameter Settings

Parameter	Value
Number of domains	2
Number of sensors per domain	20 - 100
Area size	500x500 m^2
Domain1's sink position	(125,250)
Domain2's sink position	(375,250)
Distribution of the sensors	Uniform random
Number of maximum hop	5 hops
Transmission range	100 m
Data load per packet, *b*	100 *bytes*
Path loss exponent, σ	2, 4
Number of failed nodes	4-48
Routing protocol	AODV routing

5.5.1 Discrete state vs continuous state NashQ

We evaluated two proposed methods, D-NashQ and C-NashQ, in terms of average network lifetime as shown in Figure 5.5. It can be seen that C-NashQ can achieve 23.3% longer network lifetime than D-NashQ as the network density increases. This is because D-NashQ divides battery energy which is continuous value to discrete states which may obtain suboptimal policy in the learning process. Moreover, D-NashQ requires a large number of steps to visit of each state-action pair before converging to an optimal policy resulting in slow convergence speed in the learning process as shown in Figure 5.6 because D-NashQ approaches the maximum reward approximately 1 lower than C-NashQ. If the number of steps required to visit the state-action pairs is not enough, this can result in suboptimal policies. This problem does not occur in C-NashQ because C-NashQ can learn continuous state though the feature function that is suitable for representing battery energy which is a

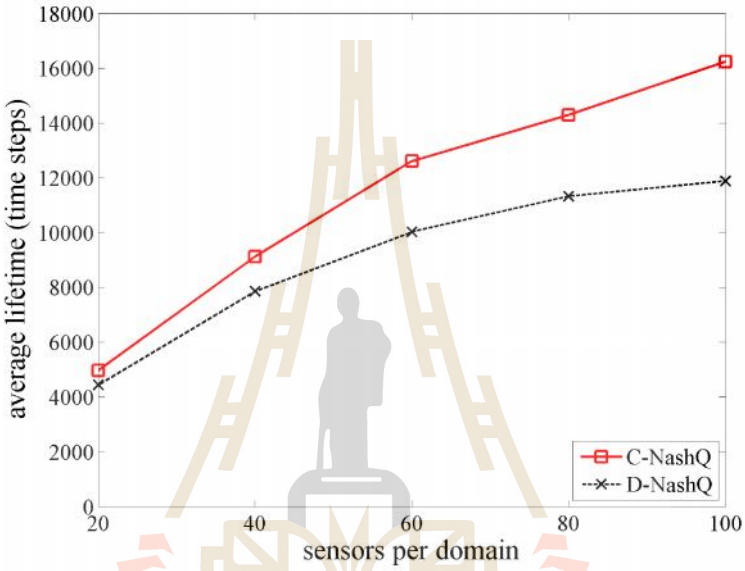


Figure 5.5 Average network lifetime

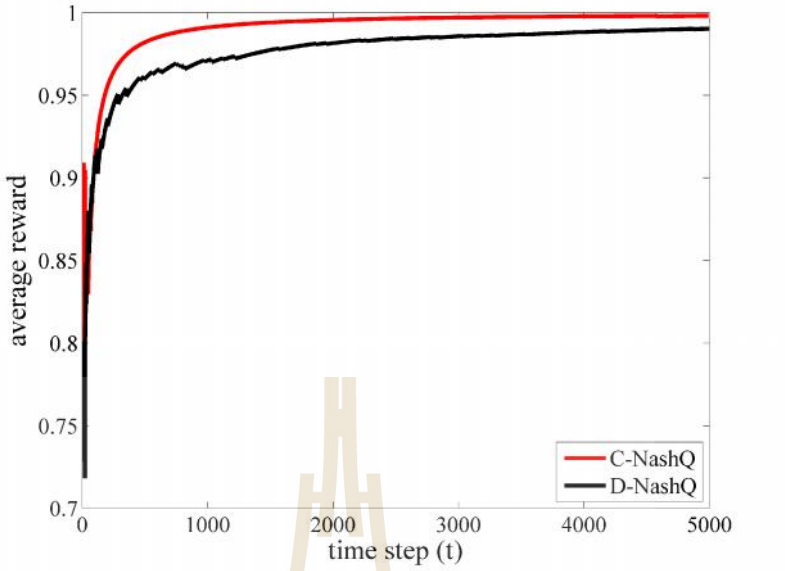


Figure 5.6 Convergence speed

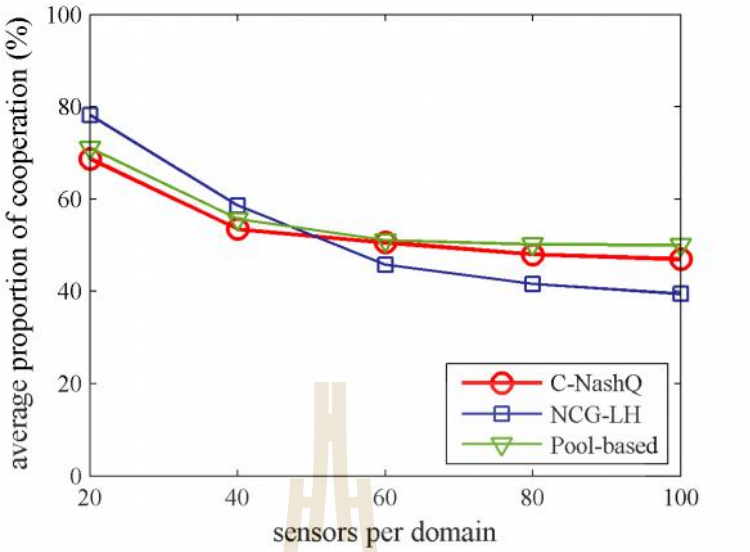


Figure 5.7 Average proportion of cooperation

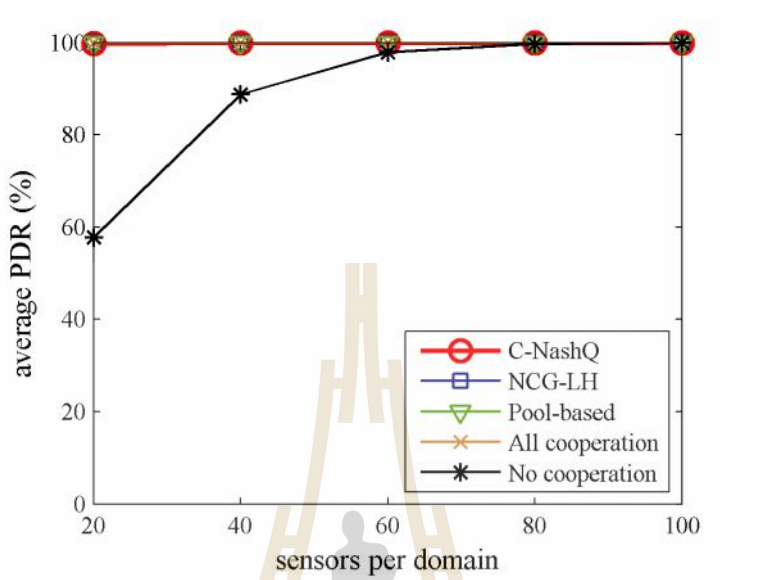


Figure 5.8 Average packet delivery ratio

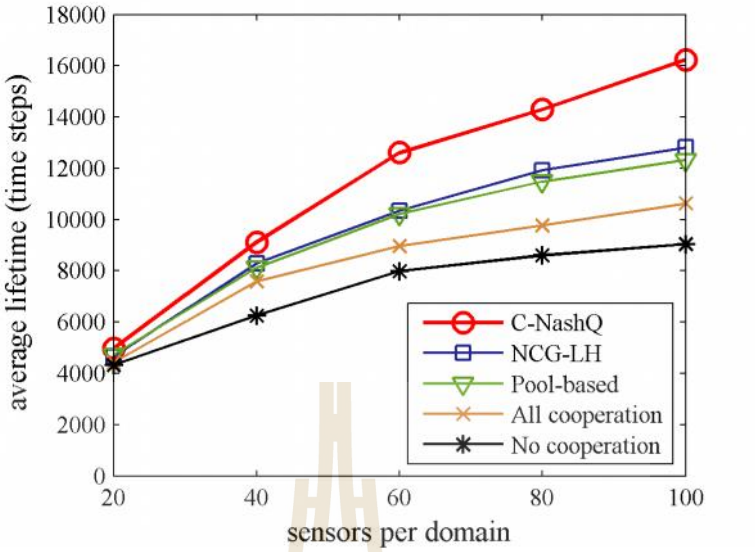


Figure 5.9 Average network lifetime

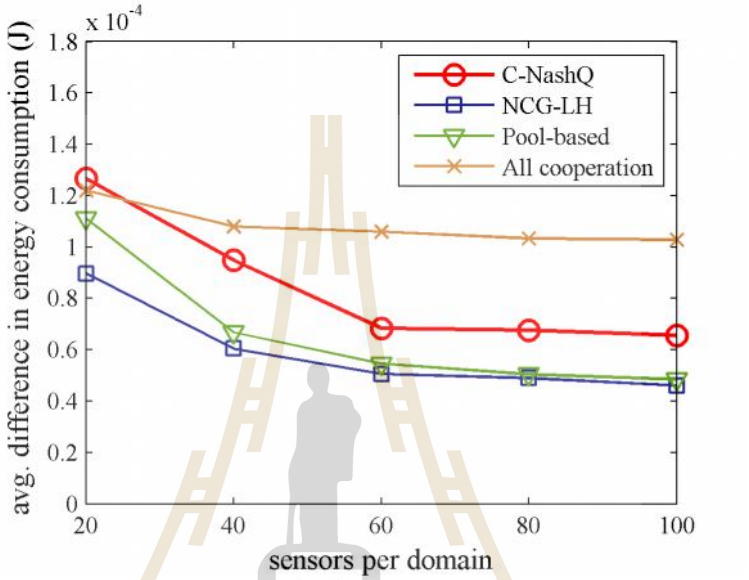
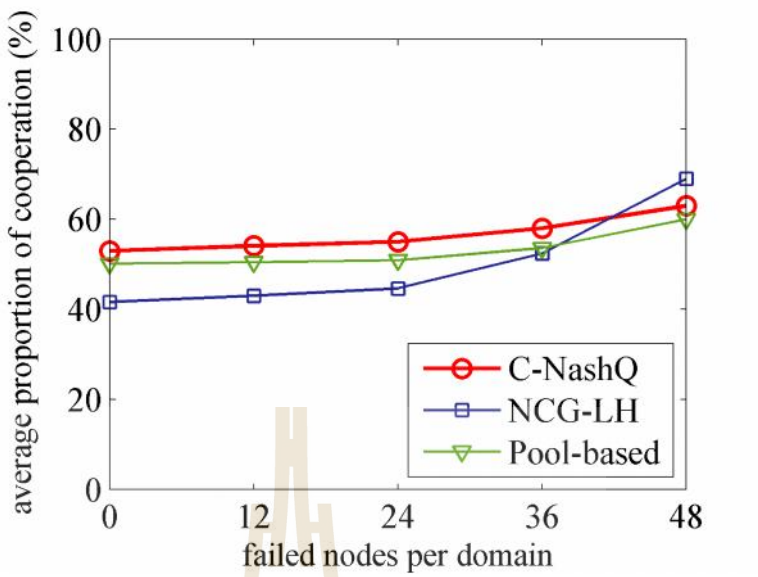
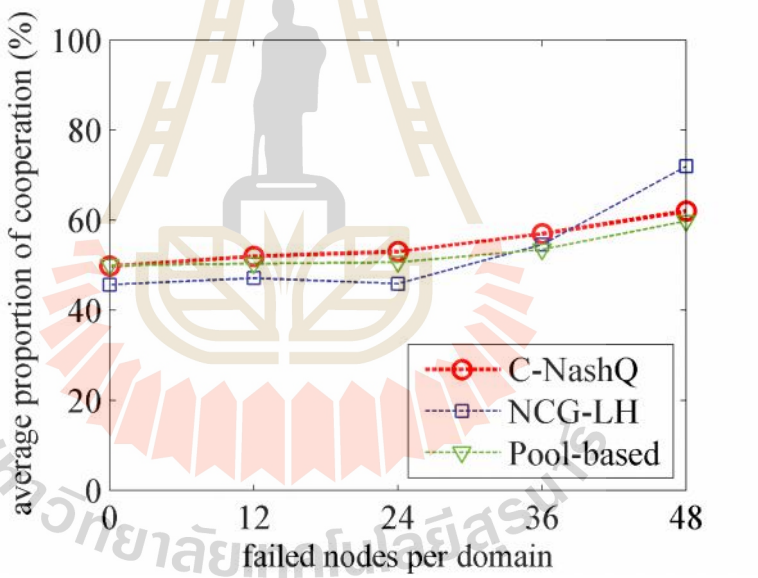


Figure 5.10 Average difference in energy consumption



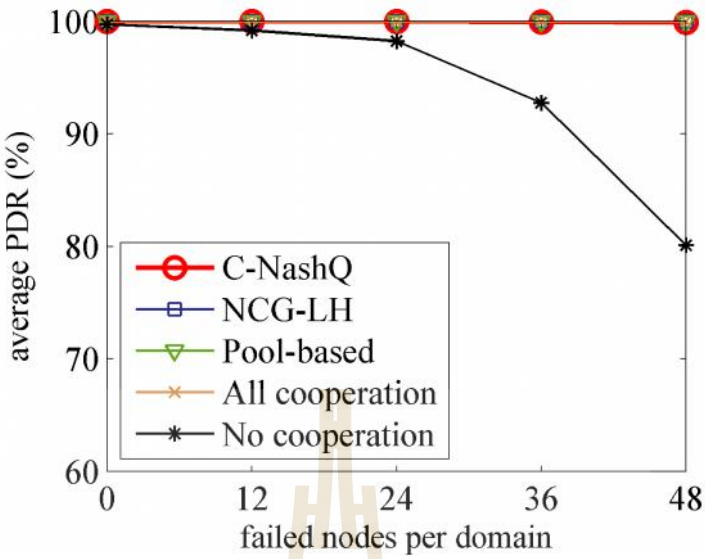
a) free space



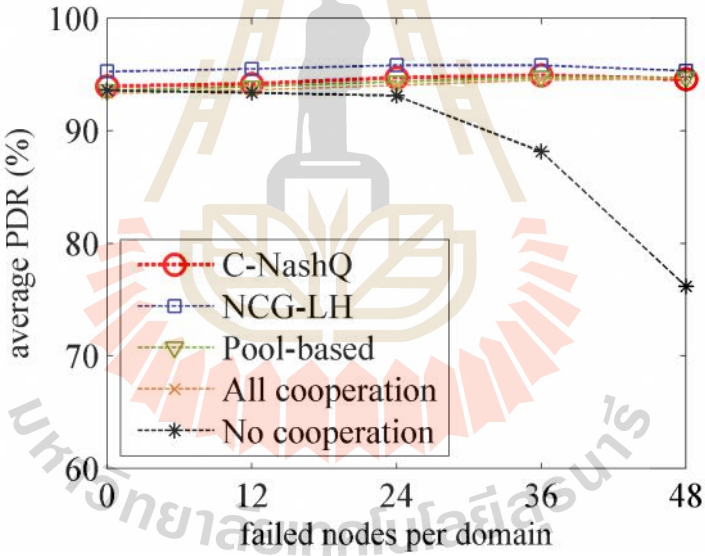
b) PLE 4

Figure 5.11 Average proportion of cooperation in various node failures under different path loss exponents

Figure 5.11 shows the average proportion of cooperation by varying number of failed nodes per domain under path loss exponent (PLE). It can be seen that C-NashQ promotes more cooperation between different domains in presence of increased failed nodes and higher PLE in order to avoid connectivity problems. Similar results are observed in NCG-LH and Pool-based routing algorithms for the same reason.



a) free space



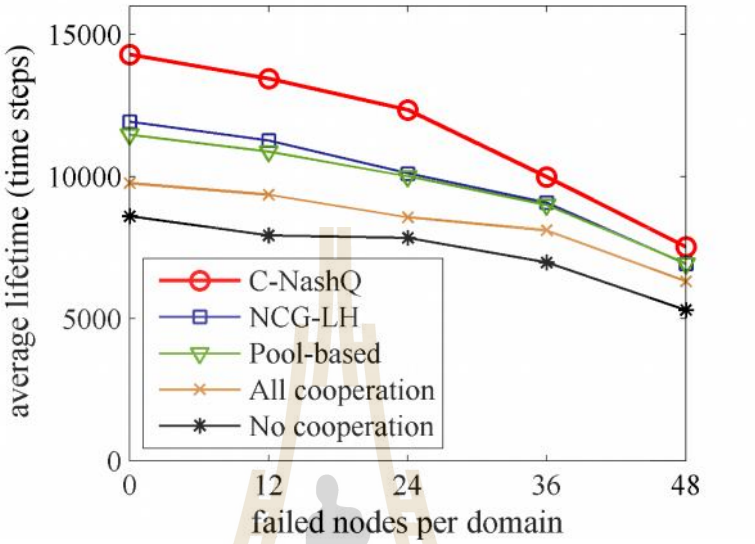
b) PLE 4

Figure 5.12 Average packet delivery ratio in various node failures under different path loss exponents

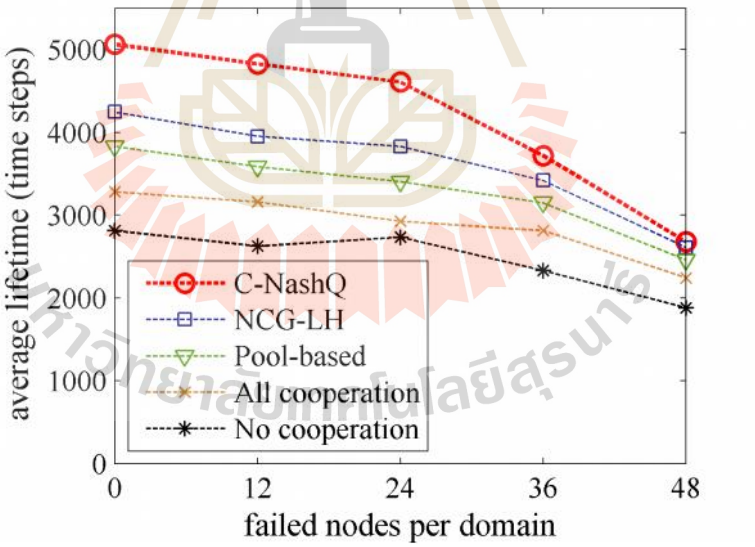
Figure 5.12 shows the average PDR of the algorithms against a varying number of failed nodes under different path loss exponents. In free space, C-NashQ can maintain 100% PDR which is comparable to NCG-LH, Pool-based and All cooperation algorithms as the number of failed nodes increases. Meanwhile, No cooperation routing algorithm can attain 80% PDR at 48 failed nodes per domain. This is because cooperation by sharing nodes between different networks can provide alternative routes for transmission of data and can thus improve the packet forwarding rate. On the contrary, No cooperation routing algorithm has the worst PDR as number of failed nodes increases. For PLE 4, C-NashQ can maintain 93% PDR comparable to NCG-LH, Pool-based and All cooperation algorithms as number of failed nodes increases to 48 nodes. On the other hand, PDR of No cooperation routing algorithm is only 75%.

Figure 5.13 depicts the average network lifetime with a varying number of failed nodes under different path loss exponents. In free space, it can be seen that C-NashQ achieves longer network lifetime than NCG-LH, Pool-based, All cooperation and No cooperation routing algorithms by 13.6%, 16.1%, 25.6% and 35.7%, respectively, on average as the number of failed nodes increases. With PLE 4, similar trends is found as in free space with C-NashQ achieving more network lifetime than NCG-LH, Pool-based, All cooperation and No cooperation routing algorithms by 12.3%, 20.6%, 29.6% and 39.8%, respectively, on average as number of failed nodes increases.

Figure 5.14 shows the average difference in energy consumption with a varying number of failed nodes under different path loss exponents. It can be seen that C-NashQ, NCG-LH and Pool-based have comparable fair energy consumption in



a) free space



b) PLE 4

Figure 5.13 Average network lifetime in various node failures under different path loss exponents

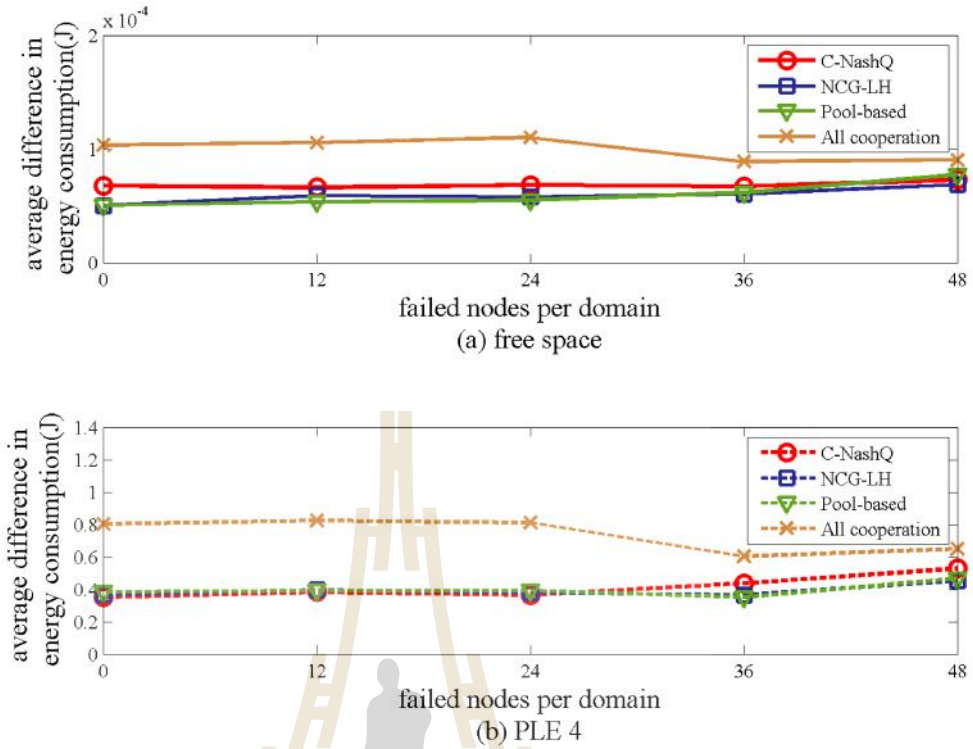


Figure 5.14 Average difference in energy consumption in various node failures under different path loss exponents

NashQ. Results show that C-NashQ can achieve 23.3% longer network lifetime than D-NashQ as the network density increases.

Moreover, this chapter also evaluated C-NashQ with four existing routing algorithms in separate sink multi-domain WSNs under uniform random topology. The results show that C-NashQ can determine suitable packet forwarding policies under various environment factors by promoting cooperation when the density of sensor is low or in presence of failed nodes and higher path loss exponents, thus improving the packet delivery ratio. Finally, C-NashQ can prolong the network lifetime by 12.3%-44.3% on average when the network density, PLE and number of failed nodes increases.

CHAPTER VI

CONCLUSION

6.1 Original contributions and findings

In multi-domain wireless sensor networks (WSNs), resource sharing and cooperation between sensor nodes belonging to different domain authorities can prolong the network lifetime and enhance reliability of packet delivery ratio. However, selfish behaviors of sensor nodes may incur in order to conserve their energy and such nodes may refuse to cooperate. However, it is possible that cooperation between sensor nodes belonging to different network authorities may not always beneficial to any WSN. Hence, the objective of this thesis is 1) to identify the parameters that effect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve; 2) to apply non-cooperative game theory to allocate packet forwarding problem in distributed multi-domain WSNs based on common sink and separate sink scenarios; 3) to obtain routing schemes which can achieve the best mutual packet forwarding strategy in non-cooperative multi-domain WSNs in a distributed manner using game theoretic reinforcement learning algorithm.

The research work carried out in this thesis is divided into three parts: the first part is designing a payoff matrix that is suitable for non-cooperative packet forwarding game. A game theory (GT) routing algorithm is proposed in Chapter 3 in order to select routes in a distributed multi-domain WSN in a *common* sink scenario. The second part is applying the GT routing algorithm to a more realistic formulation

based on a *separate* sink scenario for a packet forwarding game in multi-domain WSNs presented in Chapter 4. The last part extends the GT routing algorithm by adding a *learning* mechanism based on *game theoretic reinforcement learning*. New routing algorithms called D-NashQ and C-NashQ have been proposed in Chapter 5, which learn policies by taking the expected future payoff into consideration and can achieve suitable policy in distributed multi-domain WSNs. The original contributions in this thesis can be summarized as follows.

6.1.1 Chapter 3

The objective of this chapter is

- *To conceptually show that non-cooperative game theory can be applied to the packet forwarding problem in distributed multi-domain WSNs under the common sink scenario.*

This chapter proposes the Non-cooperative game algorithm based on Lemke Howson method (NCG-LH) algorithm to determine packet forwarding strategy between multiple domains by using Nash equilibrium (NE). The Lemke Howson (LH) method is employed to calculate the NE when pure strategy NE does not exist.

- *To study parameters that affect cooperation between multiple co-located WSNs in common sink scenario.*

Cooperation by node sharing between multi-domain WSNs may not always prolong network lifetime for any WSN. The results show that cooperation is necessary to promote when:

- low network density and without guarantee of network connectivity
- there are failed nodes that can cause failure in forwarding route path

- hostile environment in terms of higher path loss exponent (PLE)
- uniform random topology without guarantee of network connectivity

Cooperation can provide alternative routes for transmission of data, improve the packet forwarding rate and prolong the network lifetime. Moreover, the results suggest that if the networks are dense or have guaranteed of network connectivity (e.g., tree network topology), a lot of communication cost from collaboration with other networks can decrease network lifetime. Thus cooperation is not necessary in this situation.

- *To design a suitable payoff matrix for packet forwarding game in distributed multi-domain WSNs.*

A payoff matrix is proposed in NCG-LH algorithm as shown in Table 3.1 in this chapter. The proposed algorithm was compared with variations of the AODV routing protocol (i.e. the non-cooperative AODV routing and the cooperative AODV routing) in distributed multi-domain WSNs under a common sink scenario. The results show that NCG-LH obtains 12%-24% longer network lifetime than the others as network density, PLE and number of failed nodes increases in uniform random topology. Moreover, NCG-LH can achieve 20%-40% more packet delivery ratio than the non-cooperative AODV routing. Although NCG-LH performances are comparable to other algorithms in tree topology scenario with a varying network density, NCG-LH can achieve 16%-18% prolonged network lifetime and achieve 31%-37% of packet delivery ratio more than the others when subject to node failures and high path loss exponent. Finally, NCG-LH can provide fair energy consumption to all WSNs in terms of the difference in average energy consumed along a

forwarding path between both networks. The results show that NCG-LH always provide less difference of average energy consumed than the other algorithms.

The main contributions of this chapter are three-fold:

1) A non-cooperative game algorithm (NCG-LH) is proposed to distributed packet forwarding scheme in non-cooperative multi-domain WSNs based on common sink scenario.
2) Identification of parameters that affect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve, including, network density, node failure, path loss exponent, network topology and network connectivity.
3) Design of payoff matrix for non-cooperative packet forwarding game in distributed multi-domain WSNs (proposed in Table 3.1).

6.1.2 Chapter 4

The objective of this chapter is

- *To study parameters that affect cooperation between multi-domain WSNs in separate sink scenario.*

NCG-LH algorithm in Chapter 3 is evaluated in multi-domain WSNs with separate sink scenario, which is a more realistic sink scenario. The performance is compared with 3 existing algorithms include pool-based routing algorithm (Pool-based) (Kinoshita et al., 2016), which takes into account of fair energy-aware route selection in multi-domain WSNs and variations of the AODV routing protocol (i.e. the non-cooperative AODV routing and the cooperative AODV routing). The simulation results are evaluated in uniform random topology only. This is because the

proposed algorithm can distinctly provide the best performance in uniform random topology.

The results show that when the sink node of each WSN is separated, NCG-LH promotes more cooperation. Moreover, NCG-LH obtains 4.3%-31.2% longer network lifetime than the others as network density, PLE and number of failed node increases.

- *To study parameters that affect cooperation between multi-domain WSNs with varying densities in separate sink scenario*

  - The difference in node density in each domain: When number of sensors in domain 1 are denser than domain 2, NCG-LH can demote cooperation between domains due to the high availability of nodes and routes in domain N_1. This in turn, helps prolong network lifetime in domain N_2 which has less node density. The results show that NCG-LH obtains 3.3%-37.3% longer network lifetime than the others as network density, PLE and the number of failed node increases.

  - The difference of sink positions: When the sink positions are moved further away from each other, NCG-LH promotes cooperation between networks compared to the original position. NCG-LH obtains 2.6%-39.1% more network lifetime than the other algorithms as network density, PLE and number of failed node increases.

In addition, NCG-LH outperforms the other routing algorithms in terms of fair route selection by attaining the lowest average difference in energy consumption.

The main contributions of this chapter are three-fold:

1) The non-cooperative game algorithm (NCG-LH) is applied to a non-cooperative multi-domain WSNs based on separate sink scenario.

2) Investigation of fairness in terms of the difference in energy consumption between domains and comparison between a game theoretic approach (NCG-LH) and non-game theory technique (Pool-based routing)
3) Identification of parameters that effect cooperation between multiple co-located networks and fairness of benefits that the networks can achieve. These parameters include network density, node failure, path loss exponent, network connectivity, the difference of node density in each domain and sink position.

6.1.3 Chapter 5

The objective of this chapter is

- *To extend the non-cooperative game to determine long-term optimal strategies by learning from the future payoff*

In this chapter, NCG-LH is integrated with a learning mechanism i.e. by using game theoretic reinforcement learning (GTRL) in order to propose an algorithm which takes into account future (long term) benefits by allowing the agent learn strategies based on the expected future payoff (or reward). Two routing algorithms are proposed in this chapter. The first algorithm is the *discrete state Nash Q-learning* (*D-NashQ*), which is an application of the discrete state NashQ in (Hu and Wellman, 2003) to packet forwarding problem in a distributed multi-domain WSN by using payoff matrix derived in chapter 3 as a reward function. The other algorithm is the *continuous state Nash Q-learning* (*C-NashQ*) that considers the state space as continuous state, which is suitable for the continuous state of the remaining battery energy of the sensor nodes.

- *To design state space that is suitable for game theoretic reinforcement learning algorithm.*

The state space in this thesis is defined as the set of the actual battery energy of the sensor nodes. Three discrete state levels are used in D-NashQ, whereas in C-NashQ, a *feature function* is proposed for learning with a continuous state. Results show that C-NashQ can achieve 23.3% longer network lifetime than D-NashQ as the network density increases.

- *To evaluate the proposed algorithm by comparing with existing routing algorithms.*

This chapter evaluates C-NashQ with existing routing algorithms (NCG-LH, Pool-based, All cooperation and No cooperation) in separate sink multi-domain WSNs under uniform random topology. The results show that C-NashQ can determine packet forwarding policy under various environment factors and improve packet delivery ratio and prolong the network lifetime 12.3%-44.3% on average when the network density, PLE and number of failed nodes increases.

The main contributions of this chapter are four-fold:

1) Proposal of two distributed routing algorithms (D-NashQ and C-NashQ) and their application to the packet forwarding problem in multi-domain WSNs under separate sink scenario.
2) Derivation of feature function to represent the continuous state in continuous state Nash Q-learning.
3) Comparison of Nash Q-learning performance in discrete state and continuous state.

4) Performance evaluation of C-NashQ with existing routing algorithms.

6.2 Recommendation for future work

6.2.1 Extension of n-domain WSNs

In this thesis, two domain WSNs are investigated in packet forwarding problems in multi-domain WSNs. However, with the recent advancements in WSNs application in large areas such as Internet of Things (Mattern and Floerkemeier, 2010), smart grid (Fadel et al., 2015), it is possible that multiple WSNs can coexist in the same area. For this reason, routing algorithms should support resource allocation in n-domain WSNs.

6.2.2 Node/sink mobility consideration

In this thesis, the sensor nodes and sinks are assumed static. However, as an extension of WSN capabilities, the device mobility and the network dynamics provide a new chain of interesting applications such as healthcare WSNs (Lee and Chung, 2014; Shen et al., 2016), animal and agriculture monitoring (Bapat et al., 2017), vehicular WSNs (Bitam et al., 2015). In such applications, sensor may frequently encounter topology changes. Therefore, routing schemes which can efficiently locate the sensor devices, establish communication paths and determine the best mutual strategy for all agents in the multi-domain WSNs are needed.

6.2.3 Extension to heterogeneous WSNs

This thesis investigated homogenous sensor nodes so far. However, in actual WSNs may differ in many aspects such as sensor devices, battery capacity, data transmission, operation start time, and so on (Kinoshita et al., 2016; Yaqoob et al.,

2017). These factors should be taken into consideration in order to determine a suitable resource allocation policy in multi-domain heterogeneous WSNs.

6.2.4 Application to other resource allocation problems

The proposed algorithm in this thesis so far is focused on the packet forwarding problem. However, this algorithm may be applied to other resource allocation problems in multi-domain WSNs which may have limitations of energy. For instance, the problem of cluster head node selection, which is a representative node in order to send packets to base station in animal monitoring in mountain pastures (Llaria et al., 2015). Sensor devices may be set up on-body of each animal (bovines, sheep and horses) in order to know the location of each animal from each herd in mountain pastures. If each sensors belong in the same area sends the same data location to sink, it may waste energy. Therefore, the proposed algorithm can be applied to the cluster head node selection problem. Game theoretic reinforcement learning algorithm may be applied to determine a suitable cluster head node by taking the battery energy into consideration in order to prolong the network lifetime in multi-domain WSNs.

6.2.5 Testbed performance evaluation

The main objective of this thesis is to show that packet forwarding strategies in non-cooperative multi-domain WSNs can be achieved by using game theoretic reinforcement learning algorithm. Results are obtained by simulation using Visual C++ programming. Therefore, an important future direction is to extend the framework to implement in an actual sensor network testbed.

REFERENCES

Ahmed, A. A. and Fisal, N. (2008). A real-time routing protocol with load distribution in wireless sensor networks. **Journal on Computer Communications**, vol. 31, issue 14, pp. 3190–3203.

Al-Rawi, H. A., Ng M. A. and Yau, K.L.A. (2015). Application of reinforcement learning to routing in distributed wireless networks: a review. **Journal Artificial Intelligence Review**, vol. 43, issue 3, pp. 381-416.

AlSkaif, T., Zapata, M. G. and Bellalta, B. (2015). Game theory for energy efficiency in wireless sensor networks: latest trends, **Journal of Network and Computer Applications**, vol. 54, pp. 33–61.

Al-Zahrani, A.Y. and Yu, F. R. (2016). An energy-efficient resource allocation and interference management scheme in green heterogeneous networks using game theory. **IEEE Transactions on Vehicular Technology**, vol. 18, no. 7, pp. 5384-5396.

Bapat, V., Kale, P., Shinde, V., Deshpande N. and Shaligram, A. (2017). WSN application for crop protection to divert animal intrusions in the agricultural land. **Journal of Computers and Electronics in Agriculture**, vol. 133, pp. 88–96.

Bicakci, K. and Tavli, B. (2010). Prolonging network lifetime with multi-domain cooperation strategies in wireless sensor networks", **Journal on Ad Hoc Networks**. vol. 8, pp. 582-596.

Bicakci, K., Bagci, I. E., Tavli, B. and Pala, Z. (2013). Neighbor sensor networks: increasing lifetime and eliminating partitioning through cooperation. **Computer Standards Interfaces**, vol. 35, no. 4, pp. 396–402.

Bitam, S., Mellouk, A. and Zeadally, S. (2015). Bio-inspired routing algorithms survey for vehicular ad hoc networks. **IEEE Communications Surveys & Tutorials**, vol.17, issue 2, pp. 843 – 867.

Daskalakis, C., Goldberg, P. W. and Papadimitriou, C. H. (2009). The complexity of computing a nash equilibrium. **Communications of the ACM**, vol. 52, no. 2, pp. 89-97.

Debowski, B., Spachos, P. and Areibi, S. (2016). Q-learning enhanced gradient based routing for balancing energy consumption in WSNs. **IEEE International Workshop on Computer Aided Modelling and Design of Communication Links and Networks**.

Fadel, E., Gungor, V.C., Nassef, L., Akkari, N., Maik, A., Almasri S. and Akyildiz, I. F. (2015). A survey on wireless sensor networks for smart grid. **International Journal on Computer Communication**, vol. 71, pp. 22-33.

Fan, Q., Xiong, N., Zeitouni, K., Wu, Q., Vasilakos, A. and Tian, Y.C. (2016). Game balanced multi-factor multicast routing in sensor grid networks. **Information Sciences**, vol. 367–368, pp. 550–572.

Fan, Z., Tan, S. K. and Sooriyabandara, M. (2011). M2M communications in the smart grid: applications, standards, enabling technologies, and research challenges. **International Journal of Digital Multimedia Broadcasting**, vol.2011, pp.1-8.

Felegyhazi, M., Hubaux, J.P. and Buttyan, L. (2005). Cooperative packet forwarding in multi-domain sensor networks. **IEEE International Conference on Pervasive Computing and Communications Workshops**.

Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., How, J. P. (2013). A tutorial on linear function approximators for dynamic programming and reinforcement learning. **Foundations and Trends in Machine Learning**, vol. 6, no. 4, pp. 375–454.

Guizani, M., Chen, H.H. and Wang, C. (2015). **The future of wireless networks: architectures, protocols and services**, CRC Press.

Gupta, S. and Bose, R. (2015). Energy-efficient joint routing and power allocation optimisation in bit error rate constrained multihop wireless networks. **Journal on IET Communication**, vol. 9, issue 9, pp. 1174-1181.

Hu, T. and Fei, Y. (2010). QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks. **IEEE Transactions on Mobile Computing**, vol. 9, no. 6, pp. 796-809.

Hu, J., and Wellman, M.P. (2003). Nash Q-learning for general-sum stochastic games. **Journal of Machine Learning Research**. vol.4, pp.1039-1069.

Ivanov, S., Bhargava, K. and Donnelly, W. (2015). Precision farming: sensor analytics. **IEEE Intelligent Systems**, vol. 30, issue 4, pp. 76-80.

Jelicic, V., Tolic, D. and Bilas, V. (2014). Consensus-based decentralized resource sharing between co-located wireless sensor networks. **IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing**.

Jiang, T., Merrett, G. V. and Harris, N. R. (2013). Opportunistic direct interconnection between co-located wireless sensor networks. **International Conference on Computer Communication and Networks**.

Kashi, S. S. and Sharifi, M. (2013). Connectivity weakness impacts on coordination in wireless sensor and actor networks. **IEEE Communications Surveys & Tutorials**, vol. 15, issue 1, pp. 145 – 166.

Kinoshita, K., Inoue, N., Tanigawa Y. and Tode, H. (2016). Fair routing for overlapped cooperative heterogeneous wireless sensor networks", **IEEE Sensors Journal**, vol.16, issue 10, pp. 3981 – 3988.

Kulkarni, R. V., Forster, A. and Venayagamoorthy, G. K. (2011). Computational intelligence in wireless sensor networks: A survey. **Communications Surveys & Tutorials, IEEE**, vol. 13, issue 1, pp. 68-96.

Lasaulce, S. and Tembine, H. (2011). **Game Theory and Learning for Wireless Networks: Fundamentals and Applications**, Academic Press.

Lee, S. C. and Chung, W.Y. (2014). A robust wearable u-healthcare platform in wireless sensor network. **Journal of Communications and Networks**, vol. 16, issue 4, pp. 465 – 474.

Li, C., Zhang, H., Hao B. and Li, J. (2011). A survey on routing protocols for large-scale wireless sensor networks. **Journal on Sensors**, vol.11, issue 4, pp. 3498-3526.

Llaria, A., Terrasson, G., Arregui H. and Hacala, A. (2015). Geolocation and monitoring platform for extensive farming in mountain pastures", **IEEE International Conference on Industrial Technology**.

Mattern, F., and Floerkemeier, C. (2010). From the internet of computers to the internet of things. **From Active Data Management to Event-Based Systems and More**, Springer, vol. 6462, pp. 242 – 259.

Miller, D.A., Tilak, S., Fountain, T. (2005). Token equilibria in sensor networks with multiple sponsors, **International Conference on Collaborative Computing: Networking, Applications and Worksharing**, pp. 1-5.

Moosavi, H. and Bui, F. M. (2014). A game-theoretic framework for robust optimal intrusion detection in wireless sensor networks. **IEEE Transactions on Information Forensics and Security**, vol. 9, no. 9, pp.1367 – 1379.

Mulligan, G. (2010). The internet of things: here now and coming soon. **IEEE Internet Computer**, vol.14, issue.1, pp.35 – 36, 2010.

Nagata, J., Tanigawa, Y., Kinoshita, K., Tode, H. and Murakami, K. (2012). A routing method for cooperative forwarding in multiple wireless sensor networks. **International Conference on Networking and Services**.

Naruephiphat, W. and Usaha, W. (2008). Balancing tradeoffs for energy-efficient routing in MANETs based on reinforcement learning. **IEEE Vehicular Technology Conference.**

Naserian, M. and Tepe, K. (2009). Theoretic approach in routing protocol for wireless ad hoc networks, **Journal Ad Hoc Networks**. vol.7, pp. 569-578.

Niyato, D., Lu, X. and Ping, W. (2011). Machine-to-machine communications for home energy management system in smart grid. **IEEE Communication Magazines**, vol.49, issue.4, pp.53 –59.

Rashid, B., Rehmani, M. H. (2016). Applications of wireless sensor networks for urban areas a survey. **Journal of Network and Computer Applications**, vol. 60, pp. 192–219.

Rovcanin, M., De Poorter, E., Moerman I. and Demeester, P. (2014). A reinforcement learning based solution for cognitive network cooperation between co-located, heterogeneous wireless sensor networks. **Journal of Ad Hoc Networks**, vol. 17, pp. 98–113.

Shamani, M.J., Gharaee, H., Sadri, S. and Rezaei, F. (2013). Adaptive energy aware cooperation strategy in heterogeneous multi-domain sensor networks, **Procedia Computer Science**, vol. 19, pp.1047-1052.

Shen, J., Wang, C., Lai, C.F., Wang, A. and Chao, H. C. (2016). Direction density-based secure routing protocol for healthcare data in incompletely predictable networks. **IEEE Access**, vol.4, pp. 9163 - 9173.

Shoham, Y. and Brown, K.L. (2009). **Multiagent System: Algorithmic, Game-Theoretic and Logical Foundation**. Cambridge University Press.

Singhanat, K., Jiang, T., Merrett G. V. and Harris, N. R. (2015). Empirical evaluation of OI-MAC: direct interconnection between wireless sensor networks for collaborative monitoring. **IEEE International Conference on Sensors Applications Symposium**.

Singsanga, S., Hattagam, W., and Ewe, H. T. (2010). Packet forwarding in overlay wireless sensor networks using nashq reinforcement learning. **International Conference on Intelligent Sensors, Sensor Networks and Information Processing**.

Sutton, R. and Barto, A. (1998). **Reinforcement Learning: An Introduction**, The MIT Press, Massachusetts.

Vaz, P. M, Cunha, F.D., Almeida, J., Loureiro, A. and Mini, R. (2008). The problem of cooperation among different wireless sensor networks. **Proceedings of International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems**.

Wang, Y., Yu, F.R., Tang, H. and Huang, M. (2014). Mean field game theoretic approach for security enhancements in mobile ad hoc networks. **IEEE Transactions on Wireless Communications**, vol. 13, no.3, pp.1616 – 1627.

Wu, M. Y., Shu, W., (2005). InterSensorNet: strategic routing and aggregation. **IEEE Global Telecommunications Conference**.

Xu, Z., Zhu, F., Fu, Y., Liu, Q. and You, S. (2015). A Dyna-Q based multi-path load-balancing routing algorithm in wireless sensor networks. **IEEE Trustcom/BigDataSE/ISPA**.

Yang, J. and Brown, D.R. (2007). Energy efficient relaying games in cooperative wireless transmission systems. **Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers.**

Yang, J., Zhang, H., Pan, C. and Sun, W. (2013). Learning-based routing approach for direct interactions between wireless sensor network and moving vehicles. **IEEE International Conference on Intelligent Transportation Systems**.

Yaqoob, I., Hashem, I.A.T., Mehmood, Y., Gani, A., Mokhtar S. and Guizani, S. (2017). Enabling communication technologies for smart cities. **IEEE Communications Magazine**, vol. 55, issue 1. pp. 112 – 120.

Zaballos, A., Vallejo, A. and Selga, J. M. (2011). Heterogeneous communication architecture for the smart grid. **Journal on IEEE Network**, vol.25, issue.5, pp.30–37.

Ze, L. and Haiying, S. (2012). Game-theoretic analysis of cooperation incentive strategies in Mobile AdHoc. **IEEE Transactions on Mobile Computing**, vol. 11, issue 8, pp 1287-1303.

ZigBee Alliance, accessed on Dec. 28, 2015. [Online]. Available: http://www.zigbee.org

APPENDIX A

THE LEMKE-HOWSON ALGORITHM

THE LEMKE-HOWSON ALGORITHM

In this section we introduce the Lemke-Howson algorithm that finds Nash equilibrium of general sum bi-matrix game (Shoham & Brown, 2009). Example of the game is shown in Table A.1

Table A.1 Payoff matrix of sample game

A \ B	***D***	***F***
D	1 , 0	0 , 0.125
F	0 , 0.15	1 , 1

Two tableaux are required for the two agents in order to solve the game. The term r_i is the slack in the constraint $A^y \leq 1$ and s_j is the slack in the constraint $x^T B_j \leq 1$, so the following system is obtained:

$$\begin{aligned} Ay + r &= 1 \\ B^T x + s &= 1 \end{aligned} \tag{A.1}$$

Thus, the tableaux required are $r = 1 - Ay$, stated as Tableaux A and $s = 1 - B^T$ stated as Tableaux B:

Tableaux A:

$$\begin{aligned} r_1 &= 1 - y_3 \\ r_2 &= 1 \quad - y_4 \end{aligned} \tag{A.2}$$

Tableaux B:

$$\begin{aligned} s_3 &= 1 \quad - 0.15x_2 \\ s_4 &= 1 - 0.125x_1 \end{aligned} \tag{A.3}$$

The *r* terms are the *duals* of the *x*'s, while the *s*'s are the duals of the *y* terms, also known as the *slack variables* in the system.

The pivoting process start with arbitrarily choosing a variable x_i from the tableaux to bring into the basis. Then, a *minimum ratio test* determines the *slack variable* (or *dual*) to be removed by considering the coefficients of x_i, and the equation for the *slack variable* just removed is solved. The remaining equations are then solved in the chosen tableaux. The *dual* which left the basis determines the variable to enter the basis next.

Thus, starting with the variable x_1 is arbitrarily brought in, so by the *minimum ratio test*, s_4 leaves the basis, and solving s_4 for x_1 gives the following equation:

$$x_1 = 8 - 8s_4 \tag{A.4}$$

The variable x_1 is substituted into the remaining equations of Tableaux *B*, to produce:

$$\begin{aligned} s_3 &= 1 - 0.15x_2 \\ x_1 &= 8 - 8s_4 \end{aligned} \tag{A.5}$$

Since s_4 is y_4's dual, y_4 is brought in, and the pivoting process occurs once more, modifying Tableaux *A* in the process.

The procedure terminates when the initial variable chosen to enter the basis, x_i, or its *dual*, leaves. The resulting tableaux from this iterated pivoting are:

Tableaux A:

$$\begin{aligned} y_3 &= 1 - r_1 \\ y_4 &= 1 - r_2 \end{aligned} \tag{A.6}$$

Tableaux B:

$$\begin{aligned} x_2 &= 6.67 - 6.67s_3 \\ x_1 &= 8 - 8s_4 \end{aligned} \tag{A.7}$$

To achieve the NE from the tableaux, the slack variables r_i and s_i are set to 0 and the resulting values in x_i and y_i are expressed as probabilities, resulting in the final form of the NE. Thus, Eq. (A.6) becomes:

$$\begin{aligned} y_3 &= 1 \\ y_4 &= 1 \end{aligned} \tag{A.8}$$

And Eq. (A.7) becomes:

$$\begin{aligned} x_2 &= 6.67 \\ x_1 &= 8 \end{aligned} \tag{A.9}$$

Then, renormalizing the x_i and y_i to be proper probabilities,

$$NE = \left[\left(\frac{x_1}{x_1 + x_2}, \frac{x_2}{x_1 + x_2} \right), \left(\frac{y_3}{y_3 + y_4}, \frac{y_4}{y_3 + y_4} \right) \right] \tag{A.10}$$

And gets the solution

$$NE = \left[(0.545, 0.455), (0.5, 0.5) \right] \tag{A.11}$$

Or rewrite in NE policy, which is the probability over the agent's actions

$$\begin{aligned} ®_1 &= \left[0.545 \;\; 0.455 \right] \\ ®_2 &= \left[0.5 \;\; 0.5 \right] \end{aligned} \tag{A.12}$$

And can the payoff

$$\begin{aligned} NashA &= ®_1 \cdot A \cdot ®_2 \\ &= \left[0.545 \;\; 0.455 \right] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \\ &= 0.5 \end{aligned} \tag{A.13}$$

$$\begin{aligned} NashB &= ®_1 \cdot B \cdot ®_2 \\ &= \left[0.545 \;\; 0.455 \right] \begin{bmatrix} 0 & 0.125 \\ 0.15 & 0 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \\ &= 0.0682 \end{aligned} \tag{A.14}$$

APPENDIX B

EFFECT OF LEARNING PARAMETER

EFFECT OF LEARNING PARAMETER

In this section, we show how to set learning parameters that affect the NashQ algorithm proposed in Chapter 5. In NashQ algorithm, each agent has to update its new Q-value at time step *t+1* as follows :

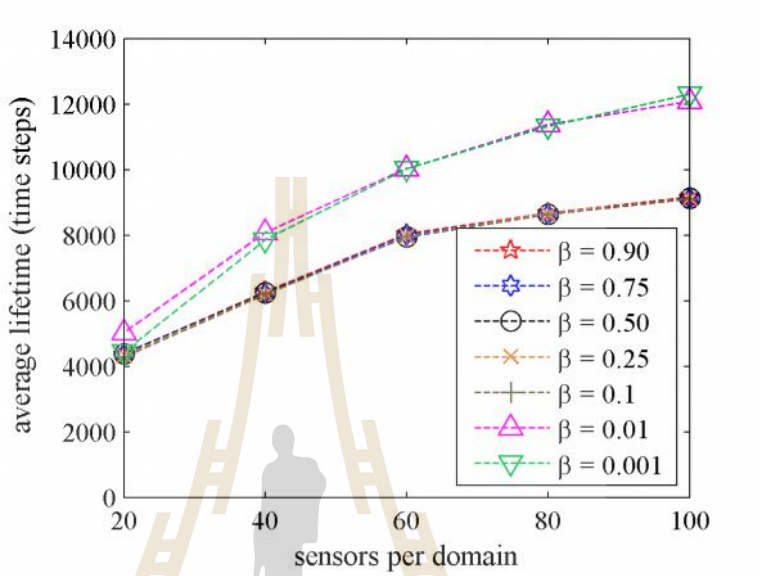
$$Q_i^{t+1}(s_i, a_1, a_2) = (1 - \sim^t) Q_i^t(s_i, a_1, a_2)) + \sim^t [r_i^t + \quad NashQ_i^t(s_i', a_1', a_2')], \qquad \text{(B.1)}$$

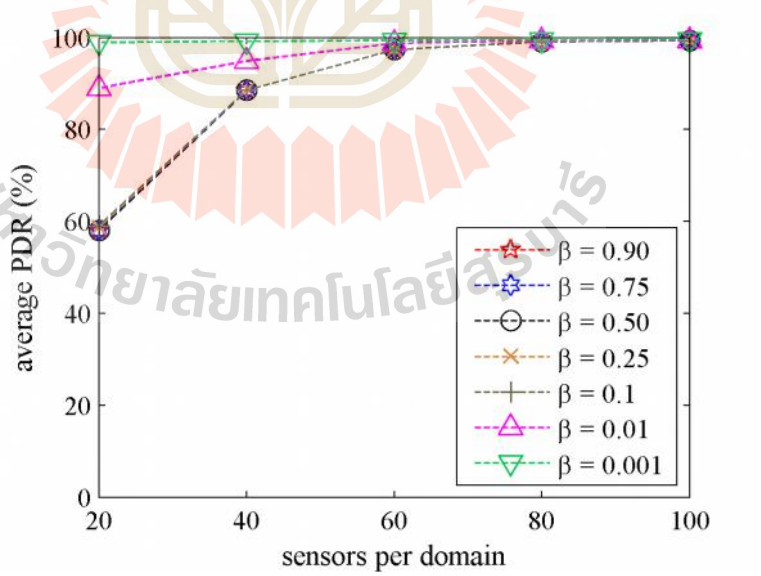
From the equation, it can be seen that the two learning parameters used in the Q-value update process are:

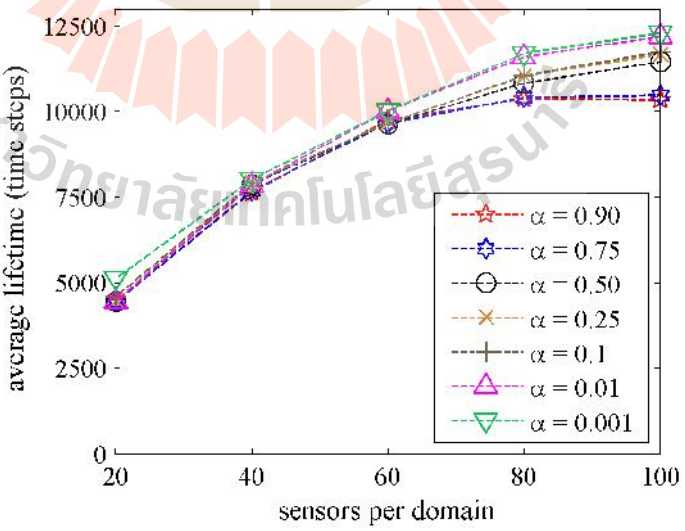
- *Discount factor*, : The discount factor defines how much expected future reward affects the immediate reward. The discount factor is usually set between 0 and 1. Setting it to 0 means that agent is interested only in the immediate reward and neglects the long term future reward. As $\to 1$, more weight is placed on the future reward in the updating process. Hence, the future reward will have more significant impact on learning the suitable cause of action.

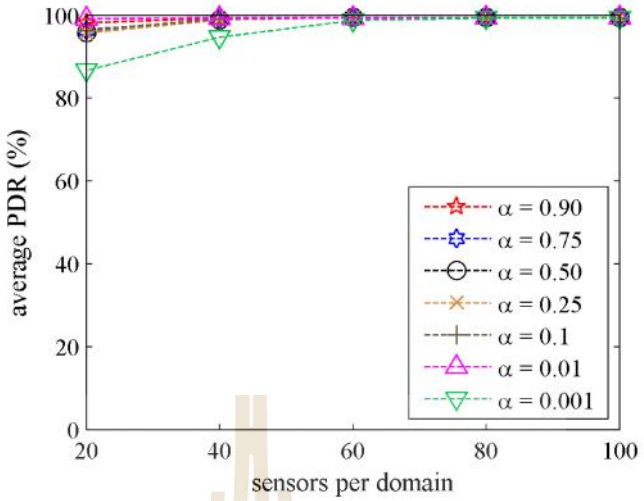
- *Learning rate*,~ : The learning rate is set between 0 and 1. It determines how fast the old Q-value is forgotten, i.e., how much weight is put on the new Q-value estimate. When ~ is 0, the Q-value will not be updated, hence nothing is learned. As ~ $\to 1$, the new Q-value estimate “forgets” the old Q-value more quickly and take the value of the new estimate $[r_i^t + \quad NashQ_i^t(s_i', a_1', a_2')]$ more rapidly as well.

To learn the optimal policy in the NashQ algorithm, we therefore study the effect of both parameters. In the experiment, we set both parameters in the range between 0-1 and observe which values that achieve the best performance in terms of network lifetime and packet delivery ratio.

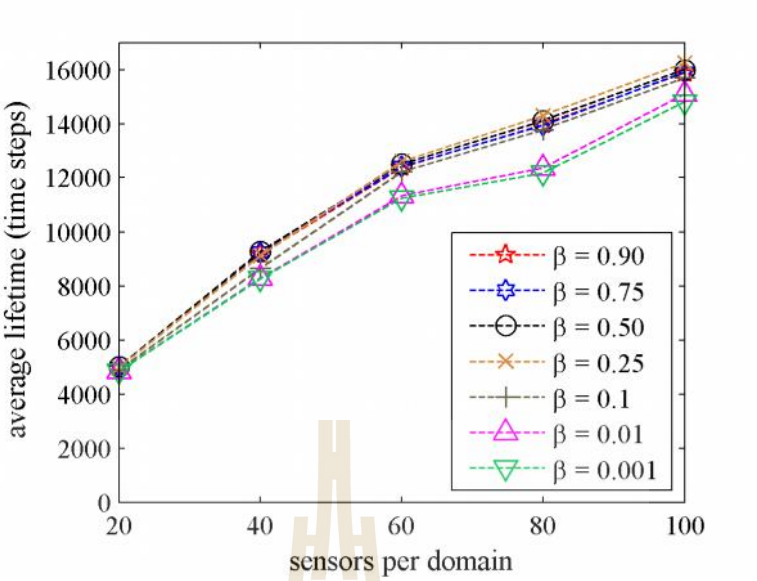


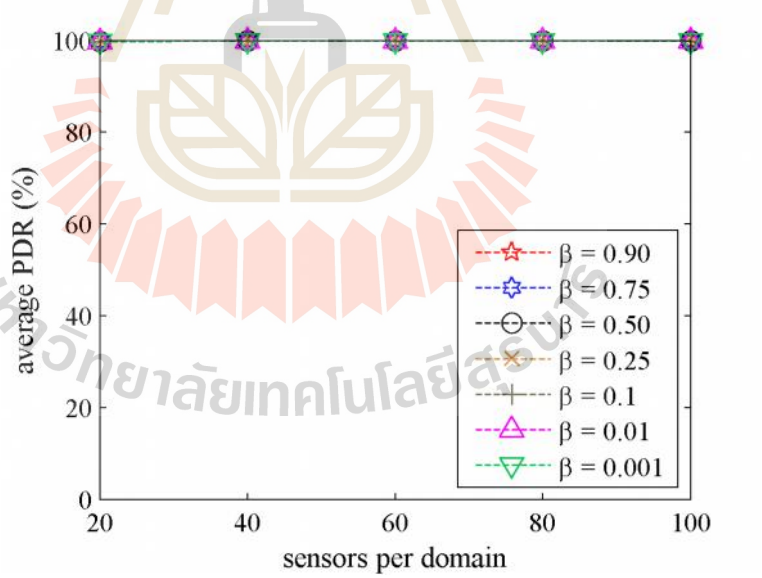


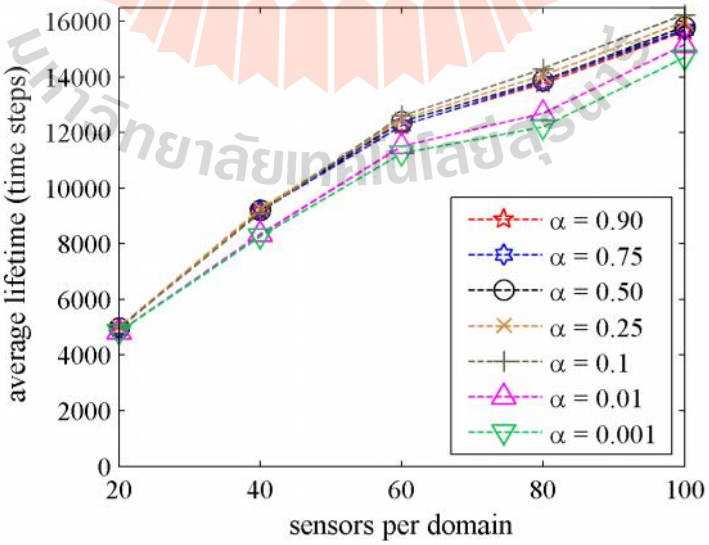

12500
10000
7500
5000
2500
0
20
40
60
80
100
avcrage lifetime (time steps)
sensors per domain
α = 0.90
α = 0.75
α = 0.50
α = 0.25
α = 0.1
α = 0.01
α = 0.001

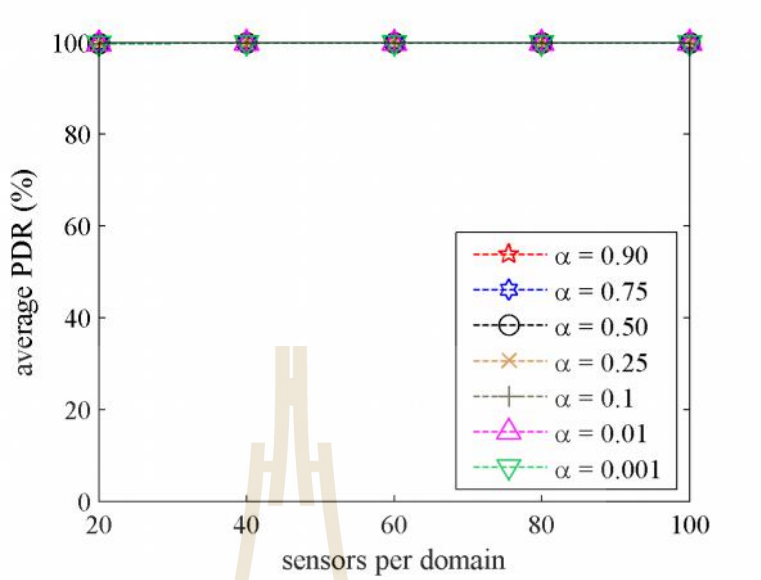


$S^n \times A^n$






average lifetime (time steps)
sensors per domain
0
2000
4000
6000
8000
10000
12000
14000
16000
20
40
60
80
100
α = 0.90
α = 0.75
α = 0.50
α = 0.25
α = 0.1
α = 0.01
α = 0.001



BIOGRAPHY

Ms. Sajee Singsanga was born on August 25, 1985 in Nakhon Ratchasima province, Thailand. She received her Bachelor's Degree and Master's Degree of Engineering in Telecommunication Engineering in 2007 and 2010, respectively, from Suranaree University of Technology. For her post-graduate, she continued to study has Master's degree in the Scool of Telecommunication Engineering Program, Institute of Engineering, Suranaree University of Technology. During Master's degree education, she was a visiting researcher at Universiti Tunku Abdul Rahman, Malaysia in a topic of resource allocation in multi-domain wireless sensor networks. She is currently pursuing her Ph.D program in Telecommunication Engineering, at the School of Telecommunication Engineering, Suranaree University of Technology. Her current research interests concern the design and simulation of network routing in wireless sensor networks.