การหาเส้นทางในเครือข่ายเซ็นเซอร์ไร้สายเคลื่อนที่สำหรับชีวการแพทย์ด้วยรีอินฟอร์สเมนท์เลิร์นนิ่ง โดยใช้ทรัสท์และเร็บพิวเทชั่น

นางสาวญานี นะพุทธะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมโทรคมนาคม

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2555

RL-BASED ROUTING IN BIOMEDICAL MOBILE WIRELESS SENSOR NETWORKS USING TRUST AND REPUTATION

Yanee Naputta

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

Suranaree University of Technology

Academic Year 2012

RL-BASED ROUTING IN BIOMEDICAL MOBILE WIRELESS SENSOR NETWORKS USING TRUST AND REPUTATION

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Master's Degree.

Thesis Examining Committee

_____

(Asst. Prof. Dr. Peerapong Uthansakul)

Chairperson

_____

(Asst. Prof. Dr. Wipawee Hattagam)

Member (Thesis Advisor)

_____

(Asst. Prof. Dr. Paramate Horkaew)

Member

_____

(Prof. Dr. Sukit Limpijumnong)

Vice Rector for Academic Affairs

_____

(Assoc. Prof. Flt. Lt. Dr. Kontorn Chamniprasart)

Dean of Institute of Engineering

ญานี นะพุทธะ : การหาเส้นทางในเครือข่ายเซ็นเซอร์ไร้สายเคลื่อนที่สำหรับชีวการแพทย์ด้วยรีอินฟอร์สเมนท์เลิร์นนิ่ง โดยใช้ทรัสท์และเร็บพิวเทชั่น (RL-BASED ROUTING IN BIOMEDICAL MOBILE WIRELESS SENSOR NETWORKS USING TRUST AND REPUTATION) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.วิภาวี หัตถกรรม, 68 หน้า.

เครือข่ายเซ็นเซอร์ทางด้านชีวการแพทย์ได้กลายเป็นกระบวนการที่มีศักยภาพในการเฝ้าระวังด้านสุขภาพของคนได้ทั้งที่บ้านและที่โรงพยาบาล การประยุกต์ใช้เซ็นเซอร์ทางด้านชีว-การแพทย์นี้เหมาะสมอย่างยิ่งสำหรับผู้สูงอายุและผู้ทุพลภาพที่ต้องการเคลื่อนไปไหนมาไหนมากกว่าถูกจำกัดให้อยู่ในสถานที่เฉพาะ เครือข่ายดังกล่าวจะช่วยให้การเฝ้าระวังสุขภาพในด้านข้อมูลทางสรีรวิทยาของผู้ป่วยเป็นไปได้อย่างต่อเนื่อง โดยเซ็นเซอร์จะถูกติดอยู่กับตัวของผู้ป่วยและส่งข้อมูลเหล่านั้นกลับไปยังศูนย์การแพทย์ และเพื่อสนับสนุนการประยุกต์ใช้งานทางด้านชีวการแพทย์นี้ พารามิเตอร์ทางด้านประสิทธิภาพของเครือข่าย เช่น อัตราความสำเร็จในการส่งแพ็คเก็ต เวลาในการส่งข้อมูลจากต้นทางไปถึงปลายทาง จะต้องเป็นไปตามความต้องการได้เพื่อให้แน่ใจว่าแพ็กเก็ตข้อมูลสามารถถูกส่งออกไปยังศูนย์การแพทย์ อย่างไรก็ตาม ในสถานการณ์ที่สมจริงมากขึ้น บางโหนดไม่ยอมให้ความร่วมมือกับโหนดอื่น เช่น ไม่ยอมส่งต่อแพ็กเก็ตที่ได้รับมาอาจเป็นเพราะแบตเตอรีหมด โหนดชำรุดหรือทำงานผิดปกติโดยไม่ทราบสาเหตุ ซึ่งจะทำให้ประสิทธิภาพของเครือข่ายลดลง

ดังนั้น วัตถุประสงค์ของงานวิจัยนี้จึงนำเสนอการปรับปรุงวิธีการหาเส้นทางในเครือข่ายเซ็นเซอร์ไร้สายเคลื่อนที่ทางด้านชีวการแพทย์โดยใช้การบูรณาการของอัลกอริธึมเรียนรู้แบบรีอินฟอร์สเมนท์ (reinforcement learning; RL) เข้ากับกระบวนการของทรัสท์และเร็บพิวเทชั่น เรียกว่า คิวอาร์ที และทำการเปรียบเทียบกับวิธีการเดิมที่มีอยู่แล้วซึ่งเรียกว่าอัลกอริธึมอาร์แอล-คิวอาร์พี (reinforcement learning based routing protocol; RL-QRP) และอัลกอริธึมที่ไม่มีการเรียนรู้เรียกว่าเทรสโฮลด์อัลกอริธึม การจำลองสถานการณ์ต่างๆถูกทดลองภายใต้เงื่อนไขของการเคลื่อนที่ของโหนด การไม่ร่วมมือของโหนด และเงื่อนไขของเวลาในการส่งแพ็กเก็ตจากต้นทางไปปลายทางที่ต้องการ งานวิจัยชิ้นนี้ได้ศึกษามาตรชี้วัดประสิทธิภาพของการหาเส้นทางสามอย่าง คือ ค่าเฉลี่ยอัตราความสำเร็จในการส่งข้อมูล (average success ratio) ค่าเฉลี่ยของเวลาในการส่งแพ็กเก็ตจากต้อนทางไปปลายทาง (average end-to-end delay) และจำนวนของเส้นทางที่พบในแต่ละความยาวของเส้นทาง (number of discovered path for each path length)

ผลการทดลองแสดงให้เห็นว่า คิวอาร์ทีอัลกอริธึมที่นำเสนอสามารถให้ประสิทธิภาพสูงกว่าอัลกอริธึมอาร์แอลคิวอาร์พีที่มีอยู่แล้วและเทรสโฮลด์อัลกอริธึมในเทอมของค่าเฉลี่ยอัตรา

ความสำเร็จในการส่งข้อมูลภายใต้เงื่อนไขของโหนดที่ไม่ให้ความร่วมมือ สูงถึง 11% และ 25% ตามลำดับ ภายใต้เงื่อนไขของโหนดที่มีการเคลื่อนที่ สูงถึง 9% และ 22% ตามลำดับ ยิ่งไปกว่านั้น ในกรณีของเงื่อนไขเวลาในการส่งแพ็กเก็ตจากต้นทางไปปลายทางที่ต้องการ คิวอาร์ทีอัลกอริธึมมีค่าเฉลี่ยอัตราความสำเร็จในการส่งของมูลมากกว่าอาร์แอล-คิวอาร์พีอัลกอริธึมถึง 11% ซึ่งจากผลการทดลองในการทดลองของเราชี้ให้เห็นว่าวิธีการทรัสท์และเร็บพิวเทชั่นสามารถนำมาประยุกต์ใช้เพื่อปรับปรุงการหาเส้นทางในเครือข่ายเซ็นเซอร์ไร้สายเคลื่อนที่ที่มีโหนดซึ่งไม่ให้ความร่วมมืออยู่ในเครือข่ายให้มีประสิทธิภาพมากขึ้นภายใต้การประยุกต์ใช้เวลาในการส่งข้อมูลจากต้นทางไปยังปลายทางที่จำกัด

สาขาวิชาวิศวกรรมโทรคมนาคม ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2555 ลายมือชื่ออาจารย์ที่ปรึกษา_____

YANEE NAPUTTA : RL-BASED ROUTING IN BIOMEDICAL MOBILE WIRELESS SENSOR NETWORKS USING TRUST AND REPUTATION. THESIS ADVISOR : ASST. PROF. WIPAWEE HATTAGAM, Ph.D., 68 PP.

MOBILE WIRELESS SENSOR NETWORKS/ REINFORCEMENT LEARNING/ TRUST AND REPUTATION/ ROUTING/NON-COOPERATIVE

Biomedical Sensor Networks have become a potential solution for monitoring health of people in their home and at hospital. Their application is especially suitable for elderly and disabled people who may prefer to be on-the-move, rather than constrained in a particular area. Such networks allow continuous monitoring of the patient's physiological information. Sensors are attached to the body and relayed back to the medical center. To support such application, network performance metrics such as packet delivery ratio, end-to-end delay must be satisfied to ensure that data packets can be routed and reliably delivered to the medical center. However, in a more realistic scenario some nodes do not cooperate with each other (i.e. by dropping packets they receive) either due to node battery depletion, malfunctioning or simply misbehaving for unknown reason thereby degrading network performance.

The underlying aim of this research is therefore to propose an enhancement to a RL-based routing in biomedical mobile wireless sensor networks by integrating it with trust and reputation, called QRT, and compare it to an existing scheme which has been used to find optimal path through experience and reward for biomedical sensor network, called reinforcement learning based routing protocol (RL-QRP) algorithm and a non-learning algorithm called the threshold. Simulations were conducted under

different mobility, malicious nodes and end-to-end delay requirement conditions. The routing performance metrics studied in this research were of average success ratio, average end-to-end delay and the number of discovered path for each path length.

The experiments results showed that proposed QRT algorithm can outperform existing RL-QRP algorithms and the threshold scheme in terms of average success ratio by up to 11% and 25%, respectively in the malicious node variation case, and up to 9% and 22%, respectively in the node mobility variation case. Furthermore, in the end-to-end delay requirement case, QRT gained 11% up to over RL-QRP algorithm. The results in our experiments suggest that trust and reputation can be applied to improve routing in presence of malicious nodes in mWSNs with stringent end-to-end delay requirements applications.

School of Telecommunication Engineering Student's Signature _____

Academic Year 2012 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 advisor, 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 express my thanks to Dr. Kae Hsiang Kwong, a senior research fellow of University of Strathclyde, Scotland, for granting me the opportunity to do research in Scotland.

I would also like to thank Asst. Prof. Dr. Peerapong Uthansakul and Asst. Prof. Dr. Paramate Horkaew for accepting to serve in my committee.

My sincere gratitude goes to the Telecommunication Research Industrial and Development Institute (TRIDI), National Telecommunication Commission Fund, Thailand for the scholarship throughout my studies and for the fruitful discussions and insights received from all the progress update meetings.

My sincere appreciation goes to Ms. Pranitta Arthans for her valuable administrative support during the course of my dissertation.

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.

Yanee Naputta

TABLE OF CONTENTS

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**

SYMBOLS AND ABBREVIATIONS

WSNs	=	Wireless sensor networks
ECG	=	Electrocardiogram
mWSN	=	Mobile wireless sensor network
MAC	=	Media access control
GPS	=	Global positioning system
RL	=	Reinforcement learning
QoS	=	Quality-of-service
RFSN	=	Reputation based framework for sensor network
RL-QRP	=	A reinforcement learning based routing protocol with QoS support for biomedical sensor networks
MDP	=	Markov decision process
C	=	Criticality of the routing device
MDP	=	Markov decision process
t	=	Time step index
α	=	Learning rate
S_t	=	State of the process at time t
S	=	State space
s	=	Current state
s'	=	Next state
A	=	Action space
a	=	Action

SYMBOLS AND ABBREVIATIONS (Continued)

$E[\cdot]$	=	Expectation operator
β	=	Discount factor
$R(s,a,s')$	=	Expected reward given any current state *s* and an action *a* with any next state *s* '
r	=	Reward
π	=	Policy
π^*	=	Optimal policy
$P[A]$	=	Distribution over the action space
$Q_t^{\pi}(s,a)$	=	The action-value function of a given policy π associates to a state-action pair (s,a) at time *t*
R_t	=	Expected discounted return of the agent at time *t*
$E^{\pi}[\cdot]$	=	Expectation operator under policy π
$V^{\pi}(s)$	=	Value function of a state (*s*) under policy π
$V^*(s)$	=	Value function of a state (*s*) under optimal policy π^*
$Q^*(s,a)$	=	The action-value function of a given optimal policy π^* associates to state-action pair (s,a)
i	=	Class of message
θ	=	Reputation value
$p(\theta)$	=	Prior distribution
$\Gamma(.)$	=	Gamma function

SYMBOLS AND ABBREVIATIONS (Continued)

D(δ)	=	Dirichlet process
δ	=	Base measure
T_{ij}	=	Trust metric
R_{ij}	=	Reputation metric
$Q(s,a)$	=	Quality of action a at state s
γ	=	Discount factor
$Q(s',a')$	=	The expectation future reward at state s' by taking action a'
Ds_i,s_{sink}	=	The distance between node s_i and destination node
Ds_j,s_{sink}	=	The distance between node s_j and destination node
Ds_i,s_j	=	The distance between node s_i and node s_j
T_Q	=	The end-to-end delay requirement
$Tdelay\ s_i,s_j$	=	The experience delay between node s_i and s_j
N	=	The number of sensor nodes
p	=	The number of success event
n	=	The number of failures event
l_{ij}	=	Level of trust at node s_j experienced by s_i
r	=	Reward function

CHAPTER I

INTRODUCTION

This chapter introduces a background on routing problems in biomedical mobile wireless sensor networks and highlights the significance of improving routing performance in such networks. It also presents the motivation for applying trust and reputation with reinforcement learning to provide a good routing solution which is the main focus of this thesis.

1.1 Significance of the Problem

A wireless sensor networks (WSN) is a network of small devices, called sensor nodes that are embedded in the real world to collect measurements of interest, e.g., humidity in the air, soil moisture, temperature of environment, pH, etc. There are numerous applications for wireless sensor networks, e.g., battlefield surveillance, medical care, wildlife monitoring and disaster response. In this research, we are interested in biomedical wireless sensor networks which measure vital sign parameters such as body temperature, blood pressure, electrocardiogram (ECG), pulse oximeters and heart rate, etc. These parameters are sensed at a patient and transmitted to a base station at a medical center. The data is used for health status monitoring, diagnosis, treatment and further analysis. For example, Varshney, (2008) and Jovanov, (2009) proposed the use of wireless sensors to monitor vital signs of patients in a hospital environment.

In medical sensor networks used for monitoring disabled/elderly patients, sensor nodes are attached to a patient's body for physiological information. In case of emergency, patients may be moved to an emergency room, or disabled/elderly patients may be on the move in the hospital, medical staff may want to know their information continuously. Therefore, a mobile wireless sensor network system (mWSN) is necessary for biomedical sensor networks. Ref. Ying Hong Wang, (2008) and Nguyen, Defago, Beuran and Shinoda (2008) conducted some initial study on the overall network lifetime in mWSNs. Mobility can further aggravate delay problems as currents paths become disconnected, new paths must be found for replacement. Most of the fundamental characteristics of mobile wireless sensor networks are the same as that of normal static WSNs. Some major differences, however, are as follows.

1) Due to the mobility, mobile WSNs have a much more dynamic topology compared to static WSNs. It is often assumed that a sink will move continuously in a random fashion, thus making the whole network dynamic.
2) It can be reasonably assumed that a gateway sink has an unlimited energy computation and storage resources. The depleted batteries of mobile sinks can be recharged or changed with fresh ones and mobile sinks have access to computational and storage devices.
3) The increased mobility in the case of mobile WSN imposes some restrictions on the already proposed routing and MAC level protocols for WSNs (Zhou, Xing, and Yu, 2006). Most of the protocols in static WSNs perform poorly in the case of mWSNs.

4) Due to the dynamic topology of mWSNs, communication links can often become unreliable. This is can be aggravated even further in hostile or remote areas where availability of constant communication channels is low.
5) Because of the mobility, location estimation plays an important role to maintain accurate knowledge of the location of the sinks or nodes. The location of the sinks or nodes can be obtained from GPS (Kim and Hong 2009 ; Yadav, Mishra, and Gore 2009 ; Kim, Lee, Yoon and Han 2009)

From the aforementioned works, the design of mobile routing is a significant and challenging field. Nowadays, there are, however, few research in routing in mWSNs. A routing technique which suitable for mWSNs. Xuedong, Balasingham, and Byun, (2008) applies reinforcement learning which is a distributive, self-adaptive, lightweight mechanism to determine paths in a hop-by-hop manner.

Reinforcement learning (RL) is a technique used to support routing in dynamic topology networks. RL is a study of how animals and artificial systems can learn to optimize their behavior by using its experience through rewards and punishments. RL algorithms have been developed to approximate solutions to sequential optimal control problems. In the standard reinforcement learning model, an agent is connected to its environment via state perception and action (Kaelbling , Littman, and Moore, 1996). There are some works which applied RL to solve routing problem in static WSNs (Karaki, and Kamal, 2004; Aghaei, Rahman, and Saddik, 2007; Forster and Murphy, 2007; 2008; wang, 2006; Dong, Agrawal, and Sivalingam, 2007). Apart from routing, some researches Seah, Tham, Srinivasan, and Xin, (2007) and Renaud, and Tham, (2006) used RL to solve coverage problems in static WSNs. Xuedong, Balasingham, and Byun, (2008) proposed a QoS routing scheme in mobile wireless

sensor networks for biomedical sensor networks. In their research, they investigated the impact of network traffic load and sensor node mobility on the network performance. However, they considered cooperative mWSNs. But as aforementioned, a more realistic scenario would require consideration of situations which some nodes do not cooperate with others. Most routing or packet forwarding schemes in the previous literature assume that nodes function properly, are trustworthy and cooperative. However, in realistic scenarios, nodes may fail to cooperate in the network due to node battery depletion, malfunctioning or simply misbehave for unknown reasons.

The most important task of biomedical sensor networks is to ensure that data delivered to the medical center or the destination node. Reputation and trust systems have proven to be useful for detecting misbehaving nodes (faulty or malicious) and for assisting the decision-making process. Reputation systems have been widely studied in the context of several diverse domain such as such as eBay (Resnick, and Zeckhauser, 2000), Yahoo auctions (Resnick et al., 2000), and Internet-based systems such as Keynote (Blaze at al., 1996), maintain reputation metrics at a centralized trusted authority. Some research designed reputation systems for ad-hoc networks i.e., Confidant (Buchegger, and Boudec, 2002) and Core (Michiardi, and Molva, 2002), etc. These systems are distributed and also maintain a statistical representation by borrowing tools from the realms of game theory. These systems try to counter selfish routing misbehavior of nodes by enforcing nodes to cooperate with each other. More recently, reputation systems were proposed in the domain of ad-hoc networks that formulate the problem based on Bayesian analytics rather than game theory (Buchegger, and Boudec, 2003a, 2003b). These systems can counter any arbitrary misbehavior of nodes. There are some works in the area of reputation and trust

systems for WSNs (Ganerial and Srivastava, 2004; Chen, 2007). Their schemes, a sensor node continuously builds a reputation value for other nodes by monitoring their behavior. Then the sensor node uses this reputation value to evaluate the trustworthiness of other nodes. Tanachaiwiwat, Dave, Bhindwale and Helmy, (2003) propose a mechanism of location-centric isolation of misbehavior and trust routing in energy constrained sensor networks. In their trust model, the trust worthiness value is derived from the capacity of the cryptography availability and packet forwarding. Ganerial and Srivastava, (2004) proposed a reputation based framework for sensor networks (RFSN) based on beliefs. Josang and Knapskog, (1998) in order to derive reputation values where each sensor node develops a reputation for each other node by making direct observations about these other nodes in the neighborhood. Reputation is represented through a Bayesian formulation, more specifically, a beta reputation system and used to help a node evaluate the trustworthiness of other sensor nodes, then, make decisions within the network. Furthermore, the statistical foundations of RFSN algorithm can be reduced to a few basic mathematical operations of addition, subtraction, multiplication and division. So, RFSN can run on resource constrained devices and available as a middleware service on Motes.

For these reasons, this research aims to handle routing in non-cooperative biomedical mWSNs using a scalable routing mechanism for mWSNs as reinforcement learning scheme and integrate with reputation and trust system for detecting and screening for malicious node behavior in mWSNs. We also study the effect of mobility, the quantity of malicious nodes and quality-of-service requirements. We finally propose a good optimal routing strategy in mWSNs which can handle mobility, malicious and end-to-end delay requirement conditions.

1.2 Research Objectives

1. To study the effects of RL algorithm on the routing performance in mWSNs.

2. To apply reputation and trust systems to solve the routing problem in mWSNs and compare with the existing routing algorithm.

3. To study the performance of QoS routing in mWSNs.

1.3 Research Hypothesis

1. RL can provide good routing solution in mWSNs.

2. Some sensor nodes are uncooperative due to various reasons such as node battery depletion, malfunctioning or simply misbehave for unknown reason.

3. Reputation and trust can avoid misbehaving nodes in mWSNs.

1.4 Basic Agreements

1. Visual C++ was used to simulate the routing protocols in mWSNs.

2. Some data in the experiments were normalized to facilitate analysis and obtain a conclusion.

1.5 Scope and Limitation

1. RL methods were studied to find a good routing strategy in mWSNs.

2. Reputation and trust were studied and applied to RL algorithm in mWSNs. Results were compared result with the existing RL-QRP algorithm.

3. Simulations were carried out by Visual C++. The experiment results were analyzed to find a suitable routing strategy for biomedical mWSNs.

1.6 Research Methodology

1.6.1 Progressions

1. Review of literature and related theories.

2. Study the existing routing methodologies in mWSNs and their performance.

3. Test the proposed reputation and trust systems with RL algorithm by simulation using Visual C++ to solve routing problems in mWSNs.

4. Analyze and conclude results.

5. Prepare publication.

6. Write thesis.

1.6.2 Research Methodology

Objective 1: To study routing problems in mWSNs.

1. Review literature and related works about routing in mWSNs.

2. Determine the advantages and disadvantages of the routing methods chosen as benchmark for this thesis.

3. Apply simulation tools such as Visual C++ to evaluate routing mWSNs under various scenarios.

4. Design experiment scenarios evaluate an existing routing algorithm (Xuedong, Balasingham, and Byun, 2008) which used a reinforcement learning method called RL-QRP to find the route.

5. Under various network scenarios, we measured the following parameters to evaluate the performance of RL-QRP in terms of the average success ratio, the average end-to-end delay and number of discovered path for each path length.

Objective 2: To apply reputation and trust systems with RL-QRP to solve the misbehaving nodes routing problem in mWSNs and compare with the original RL-QRP.

1. Survey reputation and trust methods.
2. Add malicious nodes into RL-QRP algorithm.
3. Apply the reputation and trust method to the RL-QRP algorithm.
4. Compare the results with the original RL-QRP algorithm by considering the following parameters, the average success ratio, the average end-to-end delay and number of discovered path for each path length.
5. Add QoS condition in terms of end-to-end delay requirement to the network and compare the results with original QRT and RL-QRP algorithms by considering the following parameters, the average success ratio, the average end-to-end delay and number of discovered path under different end-to-end delay requirements.

1.6.3 Research Location

1. Wireless Communication Research and Laboratory, Factory Building 4 (F4), 111 University Avenue, Muang District, Nakhon Ratchasima 30000, Thailand.

2. Centre for Dynamic Intelligent Communications (CIDCOM) within the Department of Electric and Electrical Engineering, Strathclyde University, Royal College Building, 204 George Street Glasgow G1 1XW, Scotland.

1.6.4 Research Equipments

1. Personal Computer
2. Visual C++ software

1.6.5 Data Collection

1. Information collected by reviewing literature and related works.
2. Data collected from Visual C++ simulations.

1.6.6 Data Analysis

The simulation collected data from the sensor nodes were analyzed, compared and concluded in terms of graphs and tables.

1.7 Expected Benefit

1. A suitable routing strategy for mWSNs which contain misbehaving nodes.
2. Improved routing reliability in mWSNs.

1.8 Organization of Thesis

The remainder of this thesis is organized as follows. **Chapter 2** presents the theoretical background which underlies the contribution of this thesis. Firstly, an introduction of related works followed by the introduction of Markov decision process theory, reinforcement learning (RL) and Q learning. Finally, the basic theory of reputation and trust which are integrated with the RL process to enhance routing mWSNs including malicious node is presented in this thesis.

In the first part of **Chapter 3**, we studied the existing algorithm RL-QRP and formulated of reputation and trust to evaluate the routing performance in mWSNs under various mobility and malicious nodes conditions. The proposed algorithm which integrates RL-QRP with reputation and trust called QRT and the original RL-QRP were compared in terms of the average success ratio and the average end-to-end delay. The routing performance results were evaluated and compared between the RL-QRP and QRT algorithm under different conditions of malicious node behavior, mobility and end-to-end delay requirements.

Chapter 4 summarizes all findings and original contribution in this thesis and points out possible future research directions.

CHAPTER II

BACKGROUND THEORY

2.1 Introduction

This thesis proposed a reinforcement learning based routing mechanism in biomedical mobile wireless sensor networks using trust and reputation. A wireless sensor network (WSN) is a network of small devices, called sensor nodes that are embedded in the real world to collect measurement of the interest. There are numerous applications for wireless sensor networks, e.g., battlefield surveillance, medical care, wildlife monitoring and disaster response. In this research, we are interested in biomedical wireless sensor networks to measure parameters such as body temperature, blood pressure, electrocardiogram (ECG), pulse oximeters (SpO2) and heart rate, are sensed at a patient and transmitted to a base station at a medical center. The main function of biomedical sensor networks is to ensure that data packets can be sensed and delivered to the medical center reliably and efficiently. Thus, routing protocol plays an important role in the communication stacks and has significant impact on the network performance. However, some sensor nodes may do not cooperate which each other. Nodes may drop packets they receive due to node battery depletion, malfunctioning or simply misbehave for unknown reasons. Therefore, the main focus on this thesis is to solve the routing problem for non-cooperative mWSNs based on RL by incorporating a reputation and trust mechanism.

Reinforcement learning (Sutton and Barto, 1998) is the study of how animals or machines can learn to optimize their behavior to obtain rewards and to avoid punishments. This learning scheme can permit a decision maker to learn its optimal decisions (actions) through series of trial-and-error interactions with a dynamic environment. Its main idea is to reinforce good behaviors of the decision maker while discouraging bad behaviors through a scalar reward value returned by the environment. RL relies on the assumption that the dynamics of the system satisfies a Markov decision process (MDP).

Q-learning (Watkins, 1989) is a reinforcement learning technique that approximates the optimal action-value function which is a function that gives the expected reward for taking a given action in a given state and following a fixed policy thereafter. One of the strengths of Q-learning is that it is able to compare the expected utility of the available actions without requiring a model of the environment.

Reputation and trust systems are widely used in diverse domains. E-commerce systems, such as ebay (Resnick and Zeckhauser 2000), Yahoo auctions (Resnick et al. 2000). These systems try to counter selfish routing misbehavior of nodes by enforcing nodes to cooperate with each other.

Therefore, this chapter introduces the basic theory of reputation and trust systems and theory behind reinforcement learning. It also serves as an introduction to Q-learning algorithm which is the basis of this thesis. The next section provides a background theory of Markov decision process (MDP), followed by the birth-death process, reinforcement learning (RL) and reputation and trust process. A summary is presented in the final section.

2.2 Markov Decision Process Theory

Markov decision processes (MDPs) is a model of a decision-maker interacting synchronously with the environment. Since the decision-maker sees the environment's true state, it is referred as a completely observable Markov decision process. The basis of Markov decision process is presented as follows.

2.2.1 Markov Property

Markov property refers to the memory-less property of a stochastic process. A stochastic process has the Markov property if the conditional probability distribution of future states of the process depends only upon the present state, not on the sequence of events that preceded it. A process with this property is called a *Markov process*. The Markov property states that anything that has happened so far can be summarized by the current state S_t. Therefore, 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 by;

$$P(S_{t+1} = s_{t+1} \mid S_t = s_t, ..., S_0 = s_0) = P(S_{t+1} = s_{t+1} \mid S_t = s_t). \qquad (2.1)$$

This equation is referred to as the Markov property. In other words, a stochastic process has Markov property if the probability distribution of future states of the process time t+1, given the present state at time t and all past states, depends only upon the present state and not on any past states.

2.2.2 Markov Decision Process

The probability that the process chooses *s*' as its new state is influenced by the chosen action. Specifically, it is given by the state transition probability function. Thus, the next state *s*' depends on the current state *s* and the decision maker's action *a*. But given *s* and *a*, it is conditionally independent of all previous states and actions. In other words, the state transitions of an MDP possess the *Markov property*. This state transition probability function equation is defined by;

$$P(s' \mid s, a) = P(S_{t+1} = s' \mid S_t = s, a_t = a). \tag{2.2}$$

Similarly, given any current state and action, *s* and *a*, together with any next state, *s*', the expected value of the incurred reward is;

$$R(s, a, s') = E[r_{t+1} \mid S_t = s, a_t = a, S_{t+1} = s'] \tag{2.3}$$

where $E[.]$ is the expectation operator and r_{t+1} is the reward received at time $t+1$. Equation (2.2) and (2.3), completely specify the most important aspects of the dynamics of the MDP. The simulation programming requires the exact knowledge of these two functions in order to determine the optimal policy. A MDP model can be shown in Fig. 2.1.

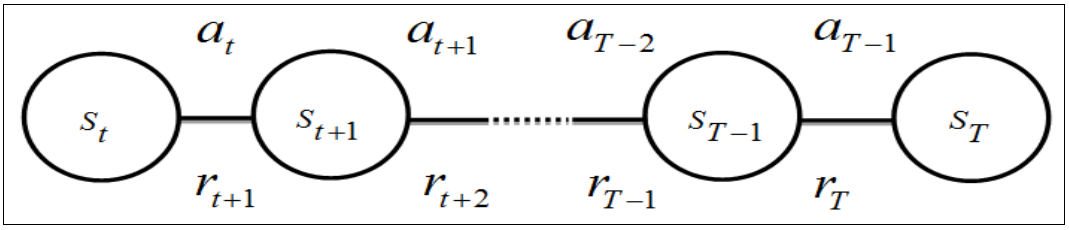

Figure 2.1 A MDP model.

A Markov decision process is a 4-tuple (*S, A, P, R*) which can describe the MDP characteristics, where *S* denotes the set of states, *A* is a finite set of actions, *P* is the probability that action *a* in state *s* at time *t* will lead to state *s′* at time *t + 1*, *R* is the immediate reward (or expected immediate reward) received after transition to state *s′* from state *s* after having taken action $a \in A$. Let $P(s' \mid s, a) \in P$ be the state transitioning model that denotes the probability of transiting to the next state $s' \in S$ after an agent takes action $a \in A$ at the current state $s \in S$.

2.2.3 Policy

A policy, π is a description of the behavior of a decision-maker, or a function mapping states to actions, $\pi: S \rightarrow A$. There are two types of policies. A *stationary policy* is a situation-action mapping, *i.e.*, it specifies an action to be taken at each state. The choice of action depends only on the state and is independent of the time step. A *non-stationary policy*, on the other hand, is a sequence of situation-action mappings, indexed by time. In this thesis, we focus on stationary policies since our data acquisition problem is based on models of sensor readings which are obtained in a particular time frame, such as in the mornings, afternoons, etc. Hence, within such period, the model maybe considered stationary hence the policy is also assumed stationary.

The objective of solving a MDP is to find a policy, π, defined as a mapping of the state space to the action space, $\pi: S \rightarrow P[A]$, where *P[A]* is the distribution over the action space. The action-value function $Q_t^{\pi}(s, a)$ of a given policy π associates a state-action pair (s, a) with an expected reward for performing action *a* in state *s* at time step *t* and policy π.

To achieve this objective, particularly in scenarios where the dynamics of the environment is difficult to model (such as in mWSNs), a technique called reinforcement learning can be used to solve MDPs.

2.3 Reinforcement Learning

Reinforcement learning (RL) is a computational approach which is concerned with how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward. In machine learning, the environment is typically formulated as a Markov decision process (MDP), and many reinforcement learning algorithms for this context are highly related to dynamic programming techniques. The main difference from these classical techniques is that reinforcement learning algorithms do not need the knowledge of the MDP and they target large MDPs where exact methods become infeasible. 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).

A reinforcement learning agent interacts with its environment in discrete time steps. At each time t, the agent receives an observation, which typically includes the reward r_t. It then chooses an action a_t from the set of actions available. The environment then moves to a new state s_{t+1} and the reward r_{t+1} associated with the transition (s_t, a_t, s_{t+1}) is determined. The goal of a reinforcement learning agent is to collect as much reward as possible. Figure 2.3 shows the agent-environment interaction in reinforcement learning.

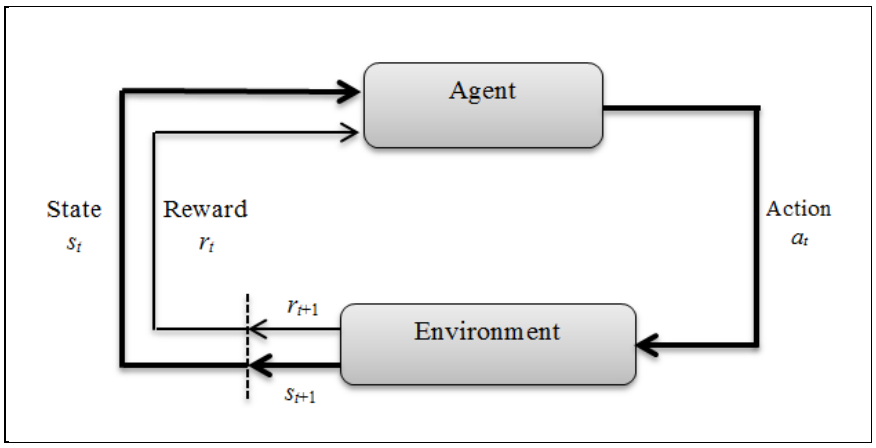

Figure 2.2 Diagram of agent-environment interaction in reinforcement learning.

2.3.1 The Value Function

Define the value function $V^{\pi}(s)$ of a policy π by;

$$V^{\pi}(s) = E^{\pi}\left[R_t \mid s_t = s\right]$$

$$= E^{\pi}\left[\sum_{k=0}^{\infty} \beta^k r_{t+k+1} \mid s_t = s\right], \tag{2.4}$$

where $R_t = r_{t+1} + \beta r_{t+2} + \beta^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \beta^k r_{t+k+1}$ is the expected discounted return of the agent, β is the discount factor which $0 \le \beta \le 1$ and $E^{\pi}[\cdot]$ is the expectation operator under policy π. Similarly, the action-value function $Q_t^{\pi}(s,a)$ of a given policy π associates a state-action pair (s,a) with an expected reward for performing action *a* in state *s* at time step *t* and following π thereafter;

$$Q_t^{\pi}(s,a) = E^{\pi}\left[R_t \mid s_t = s, a_t = a\right]$$

$$= E^{\pi}\left[\sum_{k=0}^{\infty} \beta r_{t+k+1} \mid s_t = s, a_t = a\right]. \tag{2.5}$$

2.3.2 The Optimal Value Function

Solving a reinforcement learning task means, roughly, finding a policy that achieves the maximum reward over the long run. The optimal value function denoted as $V^*(s)$ which is defined as the maximum state value function over all possible policies, at state *s*.

$$V^*(s) = \max_{\pi} V^{\pi}(s). \tag{2.6}$$

Optimal policies also share the same optimal action-value function, denoted $Q^*(s)$, and defined by;

$$Q^*(s) = \max_{\pi} Q^{\pi}(s,a). \tag{2.7}$$

The standard solution to the problem above is through an iterative search method (Puterman 1994) that searches for a fixed point of the following *Bellman* equation;

$$V^*(s) = \max_{a} \left\{ R_t + \beta \sum_{s'} P(s' \mid s,a) V^{\pi}(s') \right\}. \tag{2.8}$$

The equation (2.9) is a form of the Bellman optimality equation for $V^*(s)$. The Bellman optimality equation for $Q^*(s)$ is;

$$Q^*(s) = R_t + \beta \sum_{s'} P(s' \mid s,a) \max_{a'} Q^*(s',a'). \tag{2.9}$$

2.4 Q-learning

Q-learning is a reinforcement learning technique that works by learning an action-value function that gives the expected utility of taking a given action in a given state and following a fixed policy thereafter. One of the strengths of Q-learning is that it is able to compare the expected utility of the available actions without requiring a model of the environment. 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 of the environment and can directly approximate the optimal action-value function (Q-value) through online learning. Assume that the learning agent exists in an environment described by some set of possible states $s \in S$. It can perform any of the possible actions $a \in A$. The interaction between the agent and the environment at each instant consists of the following sequence;

- The agent senses the state $s_t \in S$.
- Based on s_t, the agent performs an action $a_t \in A$.
- As a result, the environment makes a transition to the new state $s_{t+1} = s' \in S$.
- The agent receives a real-valued reward (payoff) r_t that indicates the immediate reward value of this state-action transition.

The task of the agent is to learn a policy, $\pi : S \to A$, for selecting its next action $a_t = \pi(s_t)$ based only on the current state s_t. For a policy π, the Q-value $Q^{\pi}(s,a)$ (or state-action value) is the expected discounted cost for executing action a at state s and then following policy π thereafter. The optimal policy $\pi^*(s)$ is the policy that maximizes the total expected discount reward which received over an

infinite time. The Q-learning process tries to find $Q^*(s,a) = Q^{\pi^*}(s,a)$ in a recursive manner using available information (s_t, a_t, s', a', r_t) where s_t and s' are the states at time t and $t+1$ respectively, a_t and a' are the actions at time t and $t+1$, respectively, and r_t is the immediate reward due to a_t. The Q-learning rule at time step $t+1$ is given by;

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha\left[r_t + \beta \max_{a'} Q_t(s', a') \right] \tag{2.10}$$

where $0 \le \beta \le 1$ is a discount factor, $0 \le \alpha \le 1$ is the learning rate and $Q_t(s', a')$ is the action-value function for next state s' and next action a'.

2.4.1 Exploration

One of the most important issues for Q-learning algorithm is maintaining a balance between exploration and exploitation. Normally, the convergence theorem of Q-learning requires that all state-action pairs (s,a) are tried infinitely (Sutton and Barto, 1998). Such a balanced condition is satisfied by selecting a good action according to some probability ε and exploring new actions, otherwise. Note that ε is the probability that a greedy action is selected *i.e.*;

$$a^* = \arg\max_{\forall a \in A} Q(s,a). \tag{2.11}$$

This probability termed $\varepsilon - greedy$, significantly speeds up the convergence of the Q-value function. If the Q-value of each admissible (s,a) pair is visited infinitely often, and if the learning rate is decreased to zero in suitable way, then as

$t \to \infty$, $Q_t(s,a)$ converges to $Q^*(s,a)$ with probability 1 (Sutton and Barto, 1998). The optimal policy is defined by;

$$\pi^*(s) = \arg\max_{a \in A(s)} Q^*(s,a). \tag{2.12}$$

2.5 Trust and Reputation

In this section, we describe techniques for estimating a reputation θ based on transactional data. A transaction occurs whenever two nodes make an exchange of information or participate in collaborative process. With each exchange, the nodes generate ratings indicating the "degree of cooperation" of their partner node. For the moment, we consider reputation θ representing the probability that a given node will cooperate when asked to exchange information. Therefore, our reputations θ are contained in the unit interval [0,1], and values of θ closer to one suggest greater cooperation. In the next two section, we discuss a Bayesian framework for updating reputations given the rating from each new transaction. Within this section we address the following topics: representation of reputation update with new transactions and a trust metric as output of the reputation.

2.5.1 Representation and Update: Binary Ratings.

Suppose a transaction occurs between node *i* and *j* . Depending on the outcome, the node *i* will assign the value 1 if node *j* was cooperative and 0 otherwise. Node *i* will then update its reputation for node *j*, incorporating this new data. Independently, node *j* will create its own rating for the exchange and update its opinion of node *i*. For simplicity, we will focus on the computations carried out by node *i* with the understanding that each node in the network will perform similar operations after it completes a transaction.

Let θ denote the reputation of node *j* held by node *i*. We adopt a classical betabinomial framework for estimating reputations (Gelman et al.2003; Josang and Ismail 2002). Specifically, we assign to θ a prior distribution *p(θ)* that reflects our uncertainty about the behavior of node *j* before any transactions with *i* take place. We will take *p(θ)* from the beta family, a two-parameter class of distributions which can expressed as;

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \forall\, 0 \le \theta \le 1, \alpha \ge 0, \beta \ge 0. \tag{2.13}$$

For some choice of α and β, where $\Gamma(.)$ is the gamma function (Gelman et al. 2003). The mean of a beta distribution with parameter (α, β) is $\alpha/(\alpha+\beta)$ and its variance is $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. The beta is chosen, in part, because of its flexible and ability to peak at any value in the interval [0,1] with arbitrarily small variance (Gelman et al. 2003).

Given θ we then model our binary rating as Bernoulli observations with success probability θ. That is, let $X \in \{0,1\}$ denote node *i*'s rating of node *j* for a single transaction. Then, given *j*'s reputation θ, the probability that node *j* will be cooperative is;

$$p(X|\theta) = \theta^X(1-\theta)^{1-X}. \tag{2.14}$$

Once the transaction is complete, we update our reputation using the posterior distribution for θ;

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_{[0,1]} p(X|\theta)p(\theta)d\theta} \propto p(X|\theta)p(\theta). \tag{2.15}$$

In our case, these expressions become;

$$\theta^{X}(1-\theta)^{1-X}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+X-1}(1-\theta)^{\beta+1-X-1}, \tag{2.16}$$

which means the posterior $p(\theta|X)$ again has a beta distribution with parameters $\alpha + X$ and $\beta + 1 + X$. The utility of the choice of a beta distribution is now clear because of its relationship with the Bernoulli (binomial) distribution; the beta distribution is the conjugate prior for the bernoulli distribution. Therefore, our reputation framework requires node *i* to maintain only two parameters to describe the reputation of node *j* with very simple update rules as each new transaction occurs.

Suppose nodes *i* and *j* now conduct *n* transactions with rating $X_1, \ldots, X_n \in \{0,1\}$. Repeating the updates in the previous paragraph, we find that the posterior distribution for θ after *n* transactions is again beta with parameters updated as follows;

$$\alpha^{new} = \alpha + n\bar{X}, \quad \beta^{new} = \beta + n - n\bar{X}. \tag{2.17}$$

Therefore, after *n* transactions, the posterior mean of θ is

$$\begin{aligned}\frac{\alpha + n\bar{X}}{\alpha + \beta + n} &= \frac{\alpha}{\alpha + \beta + n} + \bar{X}\frac{n}{\alpha + \beta + n} \\ &= \frac{\alpha}{\alpha + \beta}\frac{\alpha + \beta}{\alpha + \beta + n} + \bar{X}\frac{n}{\alpha + \beta + n} \\ &= q_n\frac{\alpha}{\alpha+\beta} + (1 - q_n)\bar{X}, \end{aligned} \tag{2.18}$$

where $q_n = (\alpha + \beta)/(\alpha + \beta + n)$ is a probability that tends to zero as $n \to \infty$. This form of the updates shows clearly that we are doing a weighted average of the prior mean and the mean of the new observations. The weight on the prior mean goes to zero as the number of new observations grows very large.

2.5.2 Reputation and Update: Interval Rating.

Now we describe an update for rating that are not measured on a binary scale but instead are assigned some value in [0,1]. We can think of these rating as estimated probabilities, perhaps for the event that a particular data point exchanged between *i* and *j* is faulty. Note that the notion of estimated probabilities is much more consistent than binary ratings. In this context, we appeal to a slightly more elaborate framework involving Dirichlet processes (Ferguson 1973).

Let $D(\delta)$ be a Dirichlet process with base measure δ and let this be our prior distribution. Given observations $X_1, \ldots, X_n \in \{0,1\}$, (Ferguson 1973) tells us that posterior is again a Dirichlet process with base measure $\delta(x) + \sum_{i-1}^{n} I_{X_i}(x)$, where *I* is an indicator of a point mass at the location of the observation X_i. As we will describe in section 2.5.3, we are ultimately interested in the posterior trust, i.e. the posterior mean of the reputation distribution. When the prior mean is given by μ_δ, the posterior mean of the posterior mean of the Dirichlet process is given by;

$$q_n \mu_\delta + (1 - q_n)\overline{X}, \tag{2.19}$$

where $q_n = \delta([0,1])/(n + \delta([0,1]))$ tends to zero as $n \to \infty$ and $\mu_\delta = \int x d\delta(x)/\delta([0,1])$ is mean of the base measure. Suppose we take $\delta([0,1]) = \alpha + \beta$. Then we have;

$$q_n = \frac{\alpha+\beta}{n+\alpha+\beta}, \tag{2.20}$$

which, even though we now are dealing with real-valued observations on the interval [0,1], gives the same weights as in section 2.5.1, where we had binary cooperativeness rating. In fact, in order to match not just the weight q_n but also the prior mean μ_δ, we could take our measure δ to be $(\alpha+\beta)Beta(\alpha,\beta)$ and get exactly the same updating as in Equation 2.18 with real-valued variables $X_1, \dots, X_n \in \{0,1\}$ instead of binary variables.

Once we have seen that the update is of a generalizable form using the Dirichlet Process, we can also see that update using binary rating in section 2.5.1 can also derived within this framework. If we let the measure $\delta = \beta I_0 + \alpha I_1$, which would suggest our data are binary, then the update for the mean is again exactly Equation 2.18. We can now see that this justification is a very general one.

Following from this discussion, in order to maintain our two parameters α, β in a way so that we correctly update the posterior mean, we replace the bayesian update step with an identical bookkeeping step. After a single transaction, if the assigned probability of cooperativeness were $p \in [0,1]$, the beta parameter updates would be;

$$\alpha^{new} = \alpha + p \quad \beta^{new} = \beta + 1 - p. \tag{2.21}$$

2.5.3 Trust

The main objective of the reputation block is to expose as output metric that can be used as a representative of the subjective expectation of the other node's future behavior. Up until now we have represented *i*'s reputation of node *j*

with θ, but from here on we represent it with R_{ij} to make the pairwise reputations more explicit. Given a reputation metric R_{ij}, we define the trust metric T_{ij} as node *i*'s prediction of the expected future behavior of node *j*. T_{ij} is obtained by taking a statistical expectation of this prediction;

$$T_{ij} = E\left[R_{ij}\right] = E\left[Beta(\alpha_j, \beta_j\right] = \frac{\alpha_j}{\alpha_j + \beta_j}. \tag{2.22}$$

This trust metric can be used by a node in several ways. Some notable ones are:

(1) Data Fusion: T_{ij} can be used as a weight for a data reading reported by node *j*. The data fusion can be then performed on these weighted data readings, thereby reducing the impact of untrustworthy nodes.

(2) Node revocation: The evolution of trust over time provides an on-line tool to the end-user to detect compromised or faulty nodes. This can help the end-user to take appropriate countermeasure such as replacing the misbehaving node or sensor.

(3) Decentralized decision making: In a heterogeneous sensor network, different nodes might be equipped with different capabilities. For example, a few of them might have a more precise temperature sensor or a camera, others may be mobile, etc. Given a requirement of using a particular service from some other node in the network and faced with multiple choices, the value of T_{ij} can be used as a decision making criteria.

2.6 Summary

In this chapter, an overview of Q-learning which is a reinforcement learning method has been introduced. We provided a concise background on theories related to reinforcement learning including the Markov decision process. Furthermore, we also presented an overview of reputation and trust systems. In the next chapter a reinforcement learning based routing in biomedical mobile wireless sensor networks using trust and reputation is presented and its routing performance is compared with an existing algorithm.

CHAPTER III

RL-BASED ROUTING IN BIOMEDICAL MOBILE WIRELESS SENSOR NETWORKS USING TRUST AND REPUTATION

3.1 Introduction

In this chapter, routing issues in biomedical wireless sensor networks are investigated. Parameters such as body temperature, blood pressure heart rate are sensed at a patient and transmitted via intermediate sensor nodes to a base station at a medical center. The data is used for health status monitoring, diagnosis and treatment. For example Z. Pang, Q. Chen , and L. Zheng, (2009), E. Jovanov, C. Poon, Y. Guang-Zhong, and Y.T. Zhang, (2009) proposed the use of wireless sensors to monitor vital signs of patients in hospital and home environments.

The most important task of biomedical sensor networks is to ensure that data can be delivered to the medical center reliably and efficiently (R.S.H. Istepanian, E. Jovanov, Y.T. Zhang, 2004). Furthermore, in biomedical sensor networks, patients may be moved to an emergency room, and medical staff may want to know their information continuously. Therefore, use of a mobile wireless sensor network (mWSN) is necessary for biomedical sensors networks. A distributed, lightweight, and highly adaptive routing protocol based on methods such as reinforcement learning (RL) has been proposed for such rapidly changing wireless network conditions (E. Gelenbe and M. Gellman, 2007), (L. Xuedong, I. Balasingham, and S.S. Byun, 2008).

RL is a technique that has been used to support routing in dynamic topology networks. RL is a study of how artificial systems can learn to optimize their behavior by using its experience through rewards and punishments. There are some works which applied RL to solve routing problem in static WSNs (A. Forster, A.L. Murphy, J. Schiller, and K. Terfloth, 2008). In (E. Gelenbe and M. Gellman, 2007), the authors proposed a Cognitive Packet Network (CPN) which made routing decisions in presence of routing oscillations using RL and a neural network model. Ref. (L. Xuedong, I. Balasingham, and S.S. Byun, 2008) proposed RL-QRP, a RL-based routing protocol with routing scheme in mWSNs. They investigated the impact of network traffic load and sensor node mobility on the network performance. However, their results were based on the assumption that all nodes cooperated in the packet forwarding process. But a more realistic scenario would require consideration of situation which some nodes do not cooperate with each other (i.e., by dropping packets they receive) either due to node battery depletion, malfunctioning or simply misbehaving for unknown reason (U. Vashney, 2008). Since in biomedical sensor networks, data packets must be delivered to its destination node reliably, means to identify and avoid these malicious nodes are necessary (D. He, C. Chen, S. Chan, J. Bu, and A. Vasilakos, 2012).

Reputation and trust schemes have been used to identify well-behaved and malicious nodes in WSNs (D. He, C. Chen, S. Chan, J. Bu, and A. Vasilakos, 2012), (H. Yu, Z. Shen, C. Miao, C. Leung, and D. Niyato, 2010). In such schemes, a sensor node continuously builds a reputation value for other nodes by monitoring their behavior. Then the sensor node uses this reputation value to evaluate the trustworthiness of other nodes. Ref. D. He, C. Chen, S. Chan, J. Bu, and A. Vasilakos,

(2012) proposed a trust scheme called ReTrust for medical WSNs which is lightweight and attack-resistant. High malicious node detection rates and average packet delivery ratio were achieved via simulation and experimental test-bed. However, sensor node mobility was not explicitly addressed.

Therefore, the objective of this chapter is to solve the routing problem for non-cooperative mWSNs based on RL by incorporating a reputation and trust mechanism which screens out nodes with malicious behavior using values of reputation and trust values maintained at the sensor nodes. We compared its performance with an existing reinforcement learning routing scheme called RL-QRP (L. Xuedong, I. Balasingham, and S.S. Byun, 2008) under various mobility and malicious node scenarios.

3.2 RL-QRP

Reinforcement Learning based Routing Protocol with QoS Support for Biomedical Sensor Networks (RL-QRP) has been proposed for promote routing policies to find optimal path through experience and rewards (L. Xuedong, I. Balasingham, and S.S. Byun, 2008). They used Q-learning which learns the value of function $Q(s,a)$ to find an optimal decision policy. In each time action a is selected, the agent receives an immediate reward r from the environment. Then the agent will use this reward to update the one step rule as follows;

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a')], \tag{3.1}$$

where the Q-value, $Q(s,a)$, denotes the quality of action a at state s, α is the learning rate and γ is the discount factor. $Q(s',a')$ denotes the expectation future reward at state

s' by taking action a'. The updated Q-values then in turn affect the future decisions of the agent.

RL-QRP requires the use of location information parameters to calculate a reward following a particular action. Therefore, the protocol can find the shortest path from a beginning node to a destination node using a reward function given by;

$$r = \begin{cases} \left(\frac{\left(Ds_i,s_{sink} - Ds_j,s_{sink}\right)}{Ds_i,s_{sink}}\right) / \left(\frac{Tdelay\ s_i,s_j}{T_Q}\right), & ACK\ received \\ -\frac{Ds_i,s_j}{Ds_i,s_{sink}}, & ACK\ not\ received, \end{cases} \tag{3.2}$$

where Ds_i, s_{sink} and Ds_j, s_{sink} is the distance between node s_i, s_j and destination node s_{sink}, respectively, Ds_i, s_j is the distance between node s_i and node s_j, T_Q is the end-to-end delay requirement encapsulated in the data packet. $Tdelay\ s_i, s_j$ is the experience delay between node s_i and s_j.

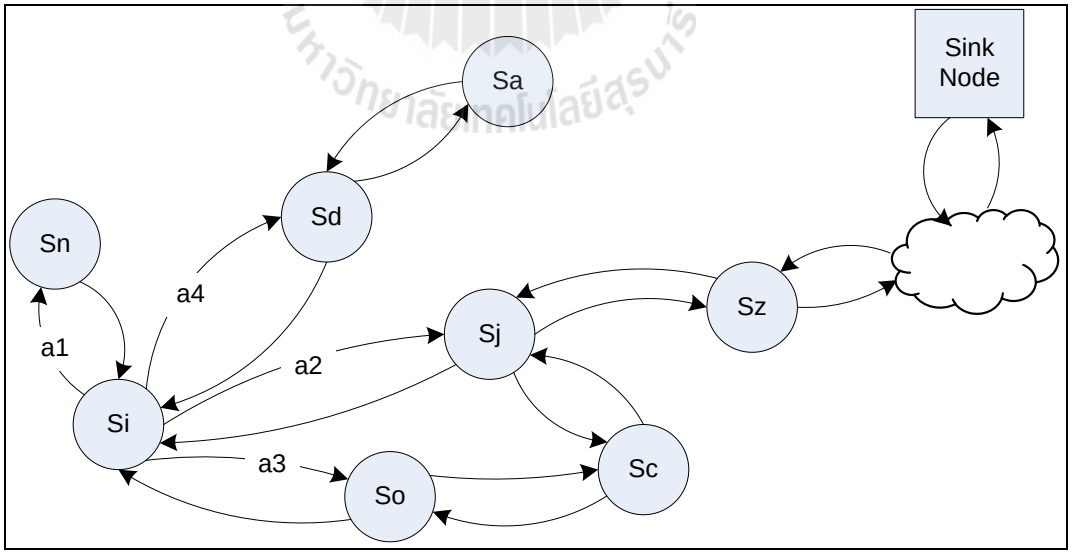

Figure 3.1 RL-QRP routing model

The basic idea of RL-QRP follows Figure 3.1. Each node in the biomedical sensor network is considered as a state belonging to set S = $\{s_i\}$, i = 1,2,…,N where N is the number of sensor nodes. For each node with a neighbor s', an action can be selected from A = $\{a(s_j|s_i)\}$. Note that $a(s_j|s_i)$ refers to a packet being forwarded from state s_i to s_j, provided that s_i and s_j are within each's other communication range. Suppose that node s_i in Figure 3.1 must forward a packet to the sink node through some intermediate node. s_i then checks the Q-value of its neighboring nodes which include $s_{n,}s_j, s_{o,}, s_d$. Then node s_i forwards the packet to the neighbor node with the highest Q-value. Suppose that s_i forwards the packet to node s_j. After that node s_i updates its Q-value $Q\left(s_i, a(s_j|s_i)\right)$ according to (3.1) with reward in (3.2). The process is repeated for node s_j and the following consecutive nodes until the packet reaches the sink node. Thus, the nodes can find the optimal route through experience and rewards without complicated prediction techniques, or explicitly frequently updating. Therefore, this process is well-suited for dynamic topologies.

3.3 Reputation

Reputation and trust systems have been proved useful mechanisms to address the threat of compromised or faulted entities. Such systems are operated by identifying selfish peers and excluding these entities from the networks. Ref. S. Buchegger and J.-Y. Le Boudec, (2002) considered routing protocols in MANETS by using both first hand and second hand information for updating reputation values. Ref. S. Ganeriwal and M. B. Srivastava, (2004) and D. He, C. Chen, S. Chan, J. Bu, and A. Vasilakos, (2012) considered both first and second hand reputation and trust-based models

developed exclusively for sensor networks. In D. He, C. Chen, S. Chan, J. Bu, and A. Vasilakos, (2012), a two-tier architecture trust management scheme was proposed in which a master node was used to compute the trust values for sensor nodes within its range. In (S. Ganeriwal and M. B. Srivastava, 2004), a watchdog mechanism was used to build their trust rating system. Given a reputation value R_{ij} obtained from the watchdog, the trust metric T_{ij} based on BETA distribution (H. Yu, Z. Shen, C. Miao, C. Leung, D. Niyato, 2010) can be computed by;

$$T_{ij} = E[R_{ij}] = \frac{p+1}{p+n+2}, \tag{3.3}$$

where T_{ij} refers to node s_i's prediction of the expected future behavior of positive outcomes of node s_j, p and n are the number of positive and negative outcomes of a specific event, respectively. R_{ij} is refer to a reputation metric. In particular, p and n are the number of successes and failures in forwarding packets between two nodes, respectively. The first hand or direct reputation value can be determined from $< p, n >$ which is the direct observation of node s_j (the observed node) experienced by node s_i. From figure 3.1, suppose that node s_i prefers to forward the data packet to the destination node by the shortest path via node s_j and s_z. In effect, an interaction occurs between node s_i and node s_j . We used a simple reputation binary rating scheme, where a successful outcome (p) is incremented if node s_j forwards the packet to node s_z and a failed outcome (n) is incremented if node s_j does not forward the packet to node s_z. Note that typically $p, n \geq 0$ so that the trust value is normalized to the range $[0,1]$, and the initial value of trust is 0.5. On the other hand, the indirect reputation value can be determined from direct reputation values of node s_j recommended by its

neighboring nodes. Although aggregated second hand information (i.e. by inquiring from watchdog the values of $< p, n >$ of other nodes which interacted with node s_j in the past) helps accelerate the calculation of the reputation value, this chapter considers the first hand observation or direct reputation for the sake of simplicity. Furthermore, drawbacks of indirect reputation include vulnerability to bad-mouthing attacks and that watchdog may not be able to capture all relevant information in the network (H. Yu, Z. Shen, C. Miao, C. Leung, D. Niyato, 2010).

3.4 RL-QRP with Reputation and Trust

In this section, RL-based routing integrated with reputation and trust, called QRT, is described. We redefine the state and action and rewards as follows:

a) Let Q($s_i, s_{j,}, l_{ij}$) denote the opinion of s_i about $s_j$which is updated when node s_j forwards or drops packets to its neighboring node;

$$Q(s_i, s_{j,}, l_{ij}) = 0.5\left\{(1-\alpha)Q(s_i, s_{j,}, l_{ij}) + \alpha\left[r + \frac{T_{ij}}{\alpha} + \left(\gamma\, max_{s'_j}\, Q(s'_i, s'_{j,}, l'_{ij})\right)\right]\right\} \quad (3.4)$$

where the Q-value, $Q(s_i, s_{j,}, l_{ij})$, denotes the quality of forwarding packets at node s_j experienced by s_i and l_{ij} denotes the level of trust at node s_j experienced by s_i which is quantized into intervals of 0.1. A trust value T_{ij} which takes values in the range [0,1] .

b) State: S = $\{s_i\}$, i = 1,2,…,N where N is the number of sensor nodes. Each node is a state in S.

c) Trust: T_{ij} is the trust value that quantifies the trustworthiness of s_j in forwarding packets from node s_i that we integtated the original Q-value of RL-QRP algorithm by average between Q-value and trust value.

d) Action: $A = \{a(s_j|s_i)\}$, s_i, $s_j \in S$. Excution of $a(s_j|s_i)$ means that the packet is forwarded from state s_i to s_j, provided that s_i and s_j are within each other's communication range.

e) Reward function: r is the reward for executing an action at node s_i (e.g. s_i forwards the packet to s_j) given by;

$$r = \left(\frac{(Ds_i,s_{sink} - Ds_j,s_{sink})}{Ds_i,s_{sink}}\right) / \left(\frac{Tdelay\ s_i,s_j}{T_Q}\right). \quad (3.5)$$

Note that we assumed that every node in the network always sends ACK back to its upstream node, regardless of their behavior. Ds_i, s_{sink} and Ds_j, s_{sink} are the distance between node s_i, s_j and the destination node s_{sink}, respectively. Ds_i, s_j is the distance between node s_i and node s_j. T_Q is the end-to-end delay requirement encapsuled in the data packet. $Tdelay\ s_i, s_j$ is the experienced delay between node s_i and s_j. The pseudo code of the proposed QRT routing algorithm is shown in Table 3.1.

TABLE 3.1 QRT routing algorithm

```
01 Begin
02 Initialization
03     Set timer for beacon exchange
04 Begin Loop
05 If timer expires
06    Broadcast beacon to immediate neighboring nodes
07    Re-set timer
08 Endif
09 If beacon packets arrives
10    Update neighboring node's position and Q-value
11 Endif
12 If data packet arrives
13          If good node
14               Random number
15                    If Random number > ε
16                         Select neighboring node with highest Q-value
17                    Else
18                         Randomly select neighboring node
19                    End if
20               Receive reward r
21               Update the Q-value
22               Update Trust
23          Else
24               Drop packets
25          End if
26 Endif
27 Go to 04
28 End
```


3.5 Performance Evaluation

In this section, we evaluated the proposed QRT routing algorithm which integrated the existing RL-QRP (L. Xuedong, I. Balasingham, and S.S. Byun, 2008) with the reputation and trust scheme. Results were compared with the original RL-QRP and a non-learning threshold reputation scheme. The latter scheme ranked the trust values of the neighboring nodes and selected the next node with the highest trust value above a predetermined threshold of 0.4 which was found to give the best performance among other threshold values. Visual C++ was used to simulate a mWSN

under various conditions according to Table 3.2 and Table 3.3. A number of nodes in the mWSN were mobile and followed the random way point mobility model which is suitable for modeling user's mobility in a confined area or within the hospital. The velocity was randomly chosen from [0,5] m/s. The remaining nodes were assumed static. These parameters are suitable for biomedical applications, where each node represents a patient who is attached with a health monitor sensor node. Each experiment was repeatedly run with different seeds, each with a runlength of 10^6 events until the sample averaged results were within a 10% range.

3.5.1. Unconstrained Traffic Demand

Initially, we evaluated the routing performance of the algorithms when there is no constrained on the QoS of the traffic demand. This experiment was divided into 2 parts where we considered the cases when the node mobility was varied and when the number of malicious nodes present in the network was varied.

3.5.1.1 Part 1 Malicious Nodes Effect

In this experiment, there are 9 mobile nodes out of 36 nodes. To study the effect of malicious nodes and the degree to which they misbehave, the number of malicious node was varied from 9 to 18 nodes and their packet dropping probability were varied from 0 to 1. The following metrics were measured:

TABLE 3.2 Simulation Parameters

Parameters	Value	
	Part 1	**Part 2**
Number of sensor nodes	36	
Node mobility	Random way point	
Pause time (s)	60	
Node velocity (m/s)	Min. 0, Max. 5	
Area size	200x200m^2	
Transmission range	50m	
Runlength (number of route requests)	10^6	
Learning rate (α) for RL-QRP, QRT	0.5	
Discount factor (γ) for RL-QRP, QRT	0.5	
Number of mobile nodes	9	0,9,18,27,36
Number of malicious nodes	9, 18	9
Probability of dropping a packet	0, 0.25,0.5,0.75, 1	0.25

- **Average success ratio (%)** is given by;

$$Average\ success\ ratio = \frac{number\ of\ discovered\ paths}{number\ of\ routing\ requests} x100\ . \quad (3.6)$$

This metric is the proportion of number of successfully discovered paths. Figure 3.2 illustrates the average success ratio for QRT, RL-QRP and threshold schemes as the packet dropping probability was varied. Note that for all packet dropping probabilities, the average success ratio of QRT was up to 11% greater than RL-QRP and up to 25% greater than the threshold scheme. Such result indicated that

QRT can identify and avoid malicious nodes more effectively than RL-QRP and threshold schemes and thereby discover more paths that can reach the destination node.

- **Average end-to-end delay:** In Figure 3.3, the average end-to-end delay is shown against the packet dropping probability. Note that the QRT showed a higher average end-to-end delay than RL-QRP. This was because QRT can discover more paths than the other schemes as shown in the previous figure. In Figure 3.4, such paths included both short paths (2, 3 hops) which was comparable to the RL-QRP, as well as long paths (4 hops up) which was discovered significantly greater than RL-QRP. The threshold scheme discovered the least number of shortest paths of all thus obtaining the highest average end-to-end delay.

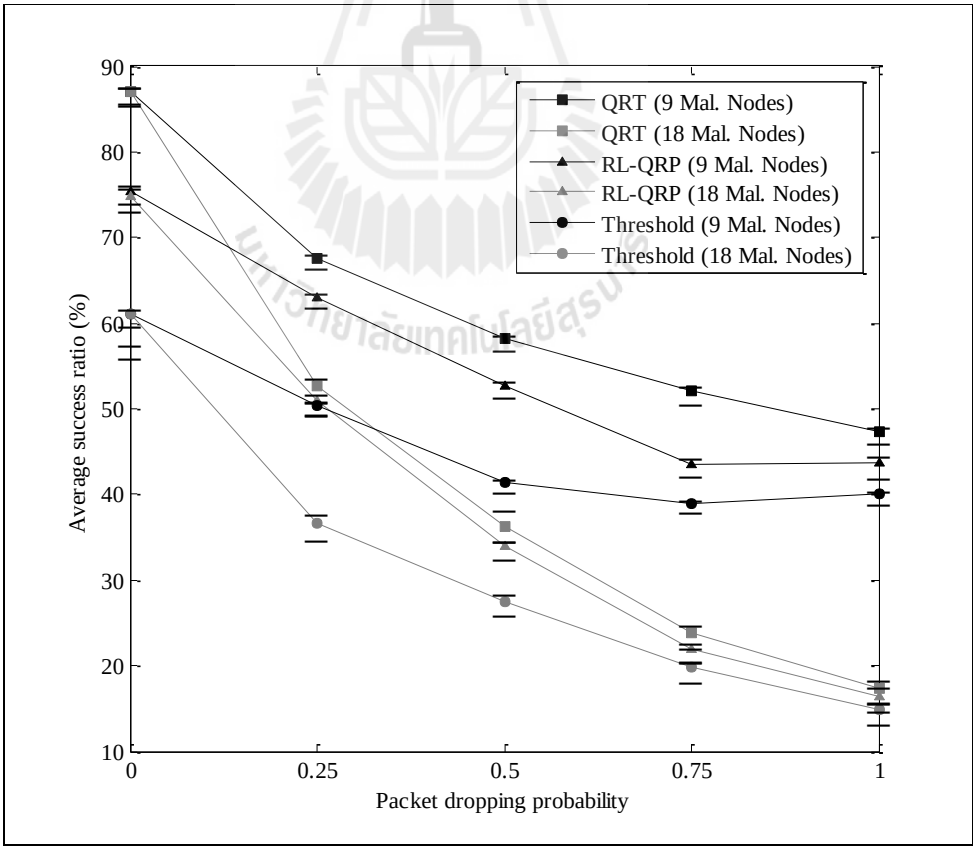

Figure 3.2 Average success ratio of discovered paths

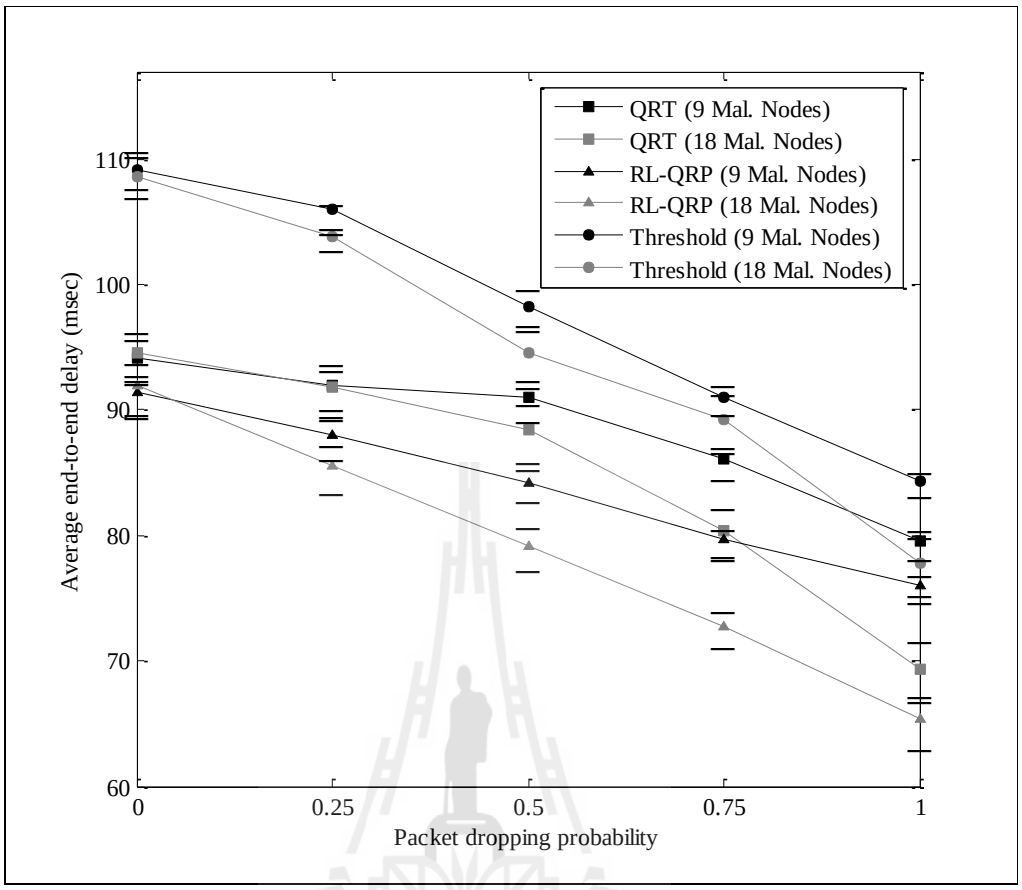

Figure 3.3 Average end-to-end delay of discovered paths

3.5.1.1 Part 2 Mobility Effect

In this part, the algorithms' performance when varying node mobility was investigated. For this scenario, 9 malicious nodes were present, each with a packet dropping probability of 0.25. Such setting was used because high success ratio were observed for all schemes. Hence, the effect from increased mobility would be more visible. The degree of mobility was varied by increasing the number of moving nodes from 0 (least mobile) to 36 (most mobile).

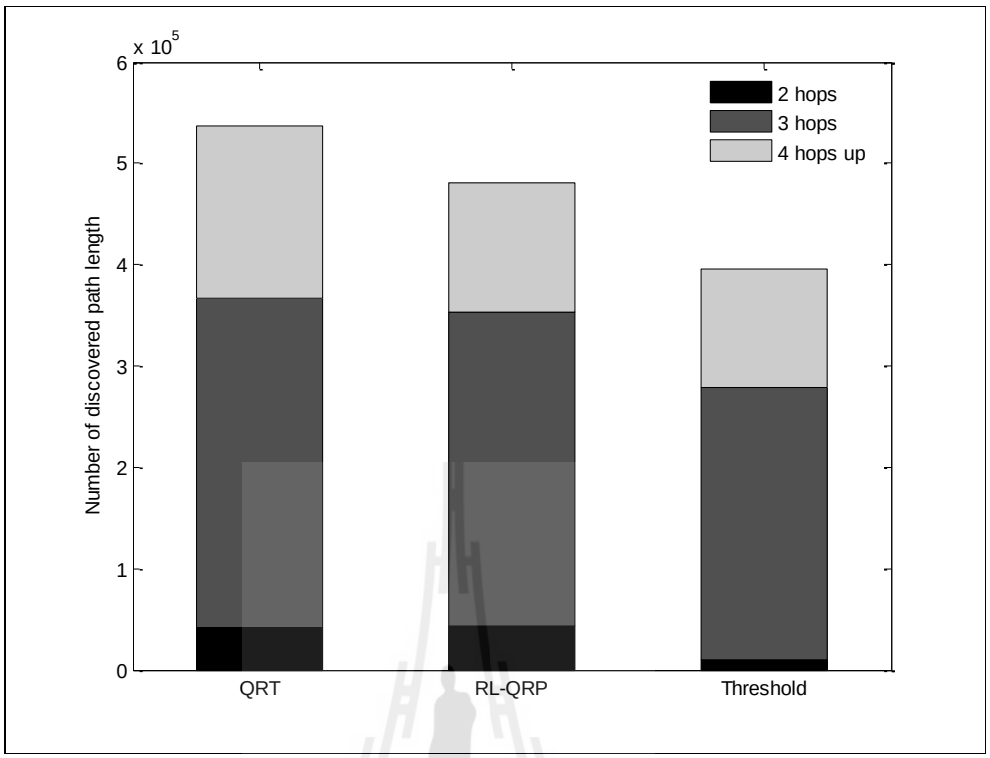

Figure 3.4 Number of discovered paths for each path length for 9 malicious nodes

- **Average success ratio (%):** Figure 3.5 illustrates the average success ratio for all schemes. Note that QRT consistently outperformed both RL-QRP and threshold schemes by up to 9% and 22%, respectively. However, the margin between QRT and RL-QRP decreased as mobility increased.

- **Average end-to-end delay:** In Figure 3.6, the average end-to-end delay is shown versus the number of moving nodes. Similar to Figure 3.3, the average end-to-end delay of QRT was greater than RL-QRP but less than the threshold scheme. This was because, in Figure 3.7, QRT can find more longer paths (4 hops up) than RL-QRP and the threshold scheme, while obtaining a comparable number of short paths (2, 3 hops) to RL-QRP. Furthermore, as the number of discovered paths gradually decreased as mobility increased, QRT consistently discovered more paths than other schemes.

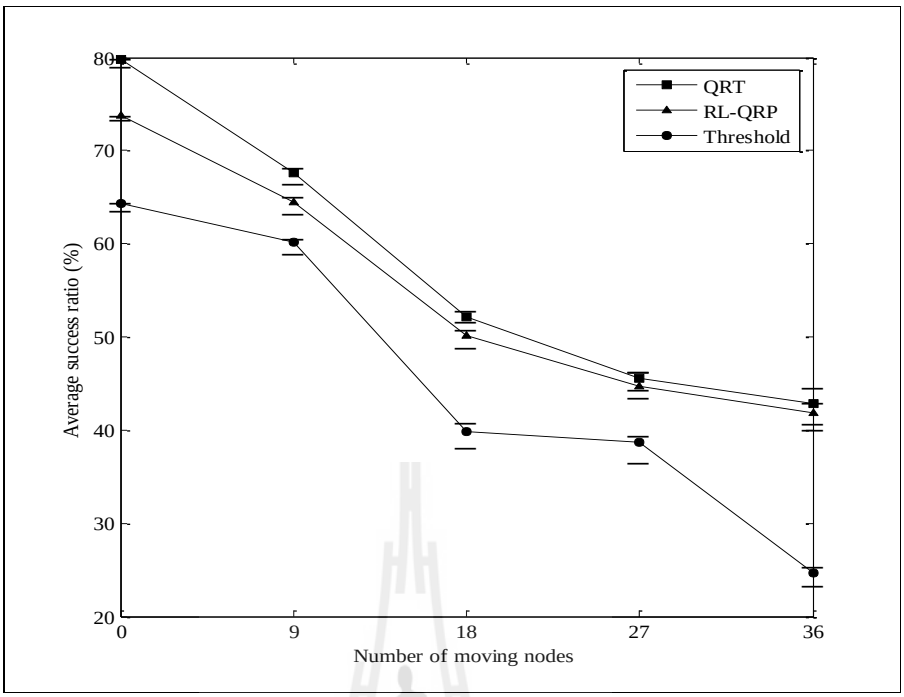

Figure 3.5 Average success ratio under various degrees of mobility

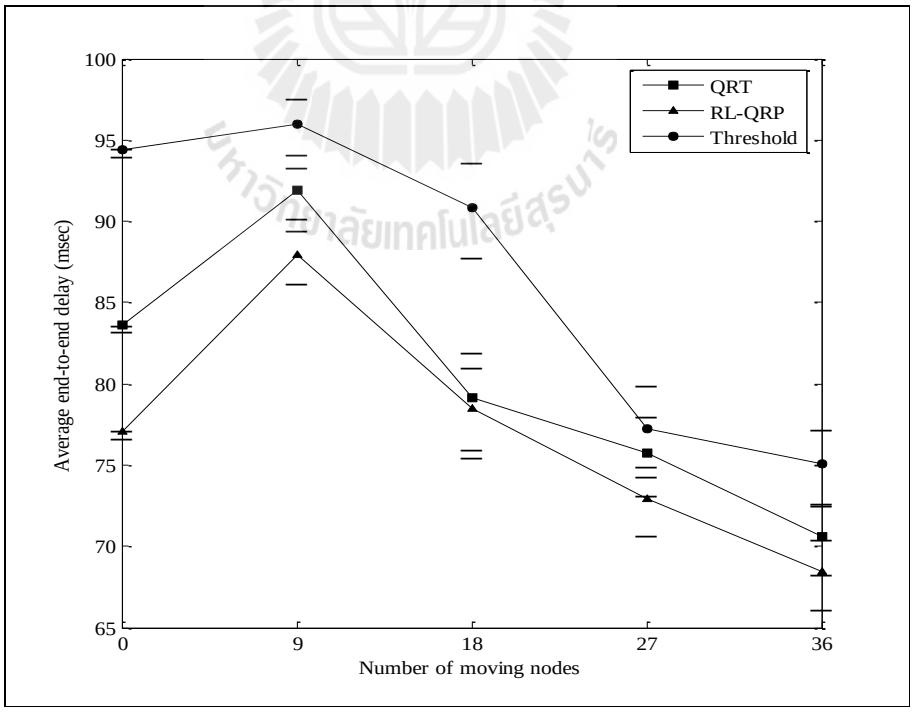

Figure 3.6 Average end-to-end delay of discovered paths

under various degrees of mobility

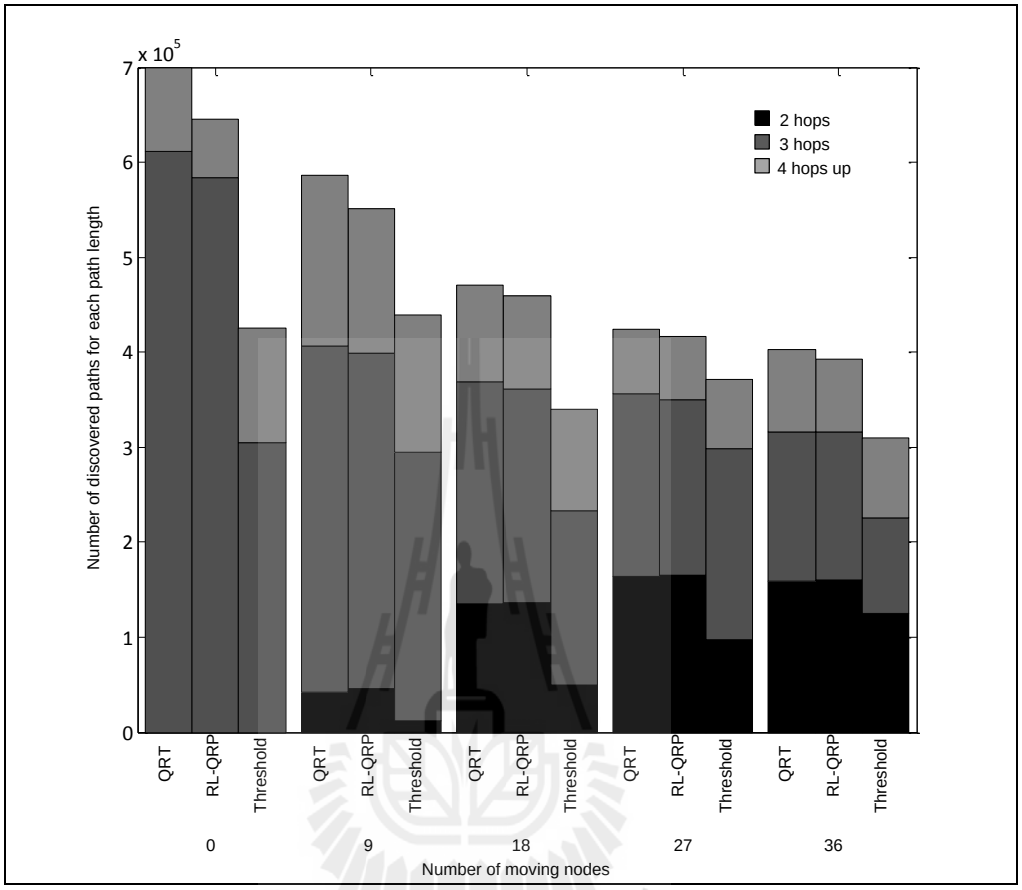

Figure 3.7 Average number of discovered path length

under various degrees of mobility

3.5.2. Traffic Demand with End-to-End Delay QoS

In this experiment, there are 9 mobile nodes and 9 malicious nodes present in the 36 node mWSN. To study the impact on the QoS on the network, the end-to-end delay requirement (T_Q) was varied to 50, 100, 200, 300 msec. The remaining simulation parameters are shown in Table 3.3.

TABLE 3.3 Simulation Parameters

Parameters	Value
Number of sensor nodes	36
Node mobility	Random way point
Pause time (s)	60
Node velocity (m/s)	Min. 0, Max. 5
Area size	200x200m^2
Transmission range	50m
Runlength (number of route requests)	10^6
Learning rate (α) for RL-QRP, QRT	0.5
Discount factor (γ) for RL-QRP, QRT	0.5
Number of mobile nodes	9
Number of malicious nodes	9
Probability of dropping a packet	0, 0.5
End-to-end delay requirement (msec)	50,100, 200, 300

- **Average success ratio**

In Figures 3.8 and 3.9, the average success ratio is shown against end-to-end delay requirement (T_Q). In this experiment, we modified the proposed QRT and the existing RL-QRP to handle different stringent end-to-end delay requirements. In particular, the reward function (r) was modified by varying T_Q accordingly for both algorithms. We thus refer to them as "QRT_T_Q reward" and "RL-QRP_T_Q reward",

respectively. Furthermore, we also evaluated a more aggressive approach in finding paths to meet the end-to-end delay requirements by allowing the agents in both algorithms to search for next hops only on paths which have the estimated delay so far not exceeding the end-to-end delay requirement. Such modification discovers paths which strictly satisfy the QoS requirement, therefore we refer to them as "QRT_strict" and "RL-QRP_strict", respectively. The value of T_Q was varied in the range 50-300 msec. We considered the cases when the packet dropping probability were 0 and 0.5. From Figures 3.8 and 3.9, we can see that in QRT consistently outperform RL-QRP. In addition, the average success ratio "QRT_T_Q reward" and "RL-QRP_T_Q reward" are greater than "QRT_strict" and "RL-QRP_strict". The reason was because "QRT _T_Q reward" and "RLQRP_T_Q reward" cannot screen out the paths whose path delay exceed the end-to-end delay requirement as shown in Figures 3.12-3.15. Furthermore, the average success ratio of "QRT_strict" and "RL-QRP_strict" decreased as T_Q became more stringent because these two methods conservatively filter out paths that have delay more than T_Q.

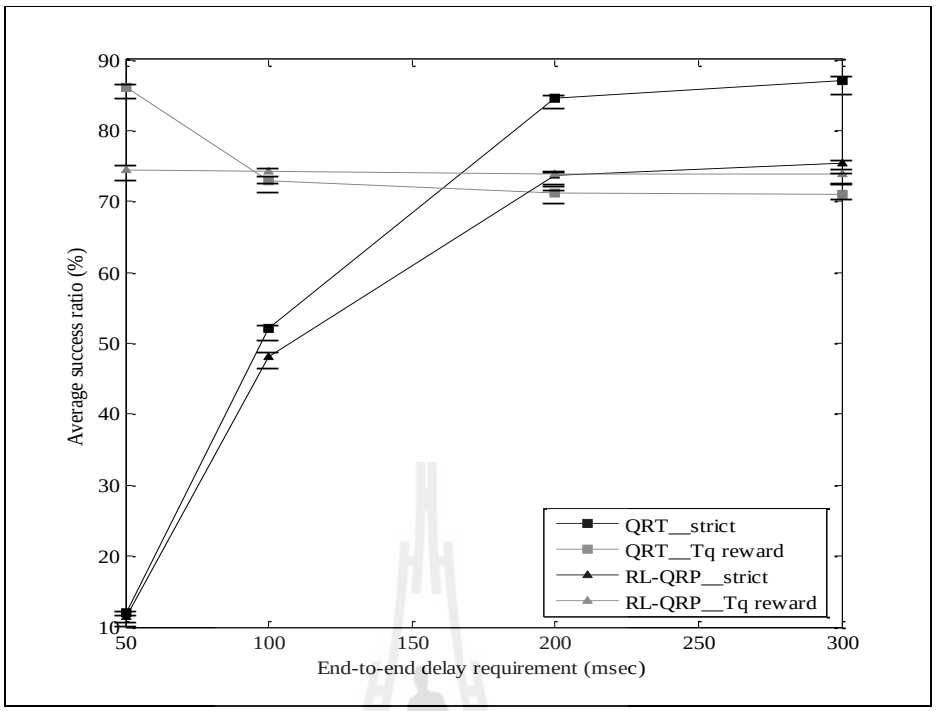

Figure 3.8 Average success ratio under different end-to-end delay requirements and 0 probability of malicious node

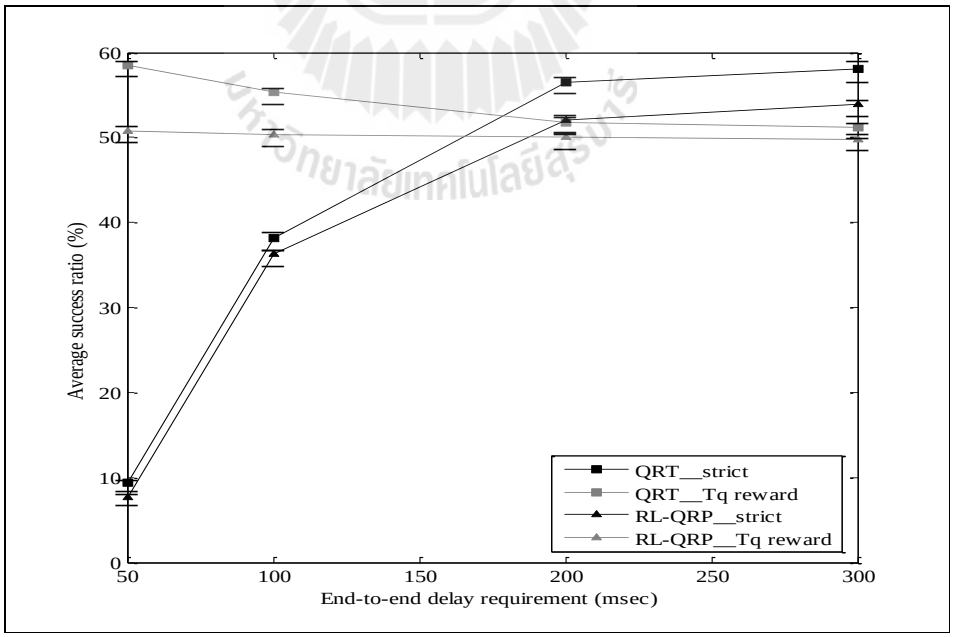

Figure 3.9 Average success ratio under different end-to-end delay requirements and 0.5 probability of malicious node

- **Average end-to-end delay**

In Figures 3.10 and 3.11, the average end-to-end delay is shown against the end-to-end delay requirement when the packet dropping probability is 0 and 0.5, respectively. Note that the average end-to-end delay of QRT and RL-QRP are similar. The average end-to-end delay of "QRT_strict" and" RL-QRP_strict" strictly satisfy T_Q because these schemes select only paths whose delays are not over T_Q. However, "QRT_T_Q reward" and "RL-QRP_T_Q reward" cannot screen out such paths delays by means of reward modification alone.

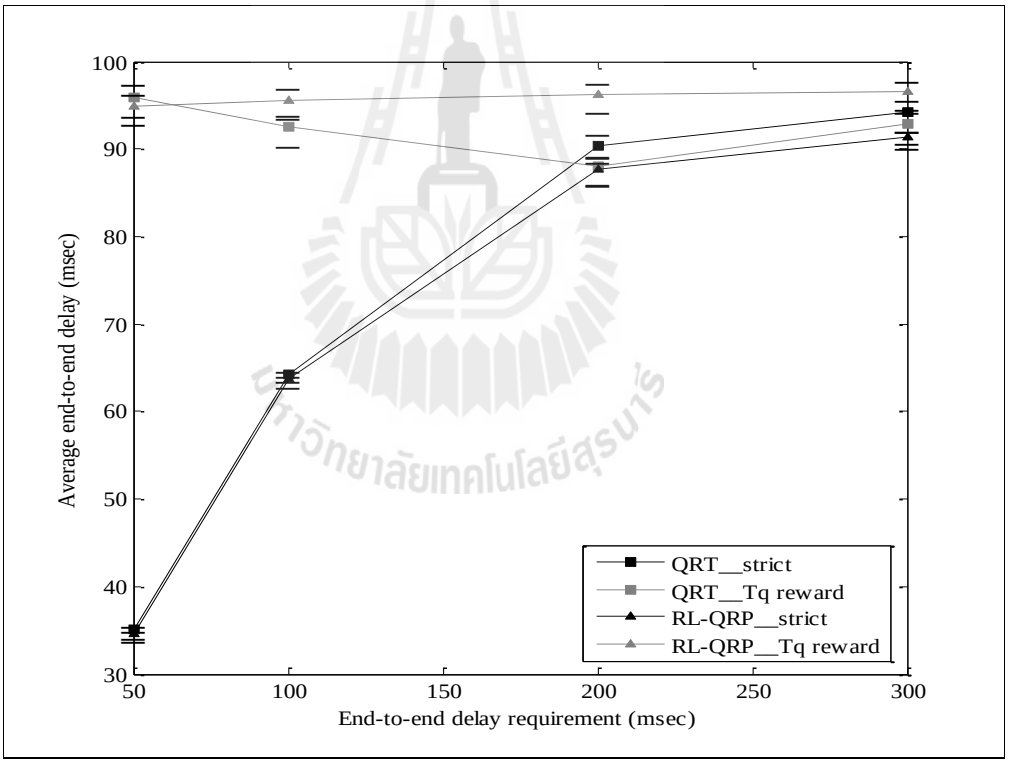

Figure 3.10 Average end-to-end delay under different end-to-end delay requirements and 0 probability of malicious node

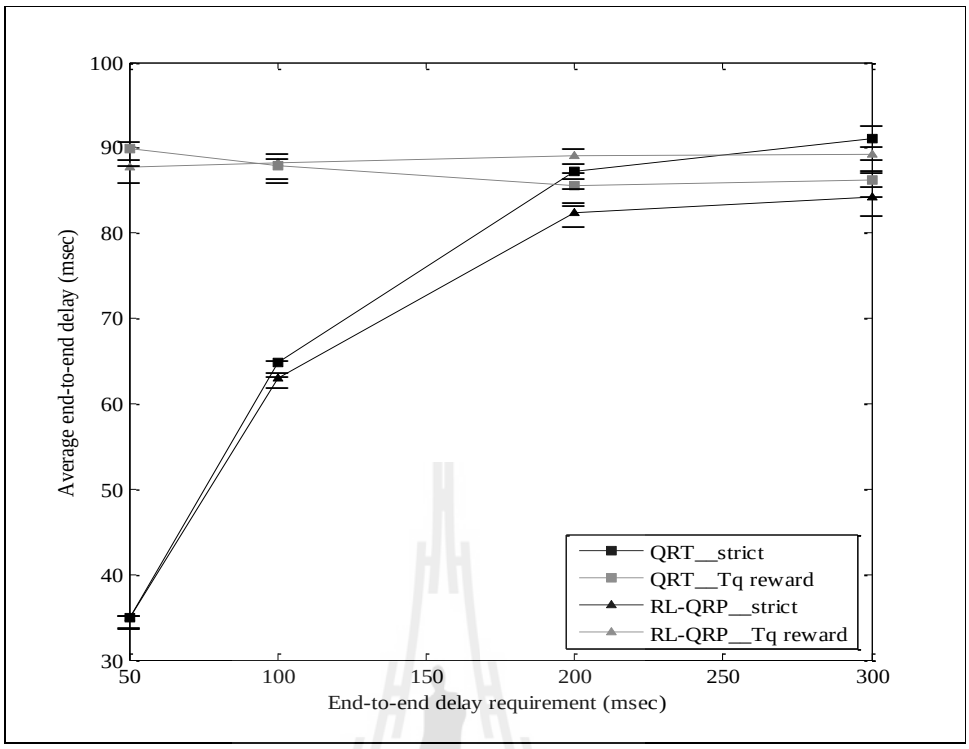

Figure 3.11 Average end-to-end delay under different end-to-end delay requirements and 0.5 probability of malicious node

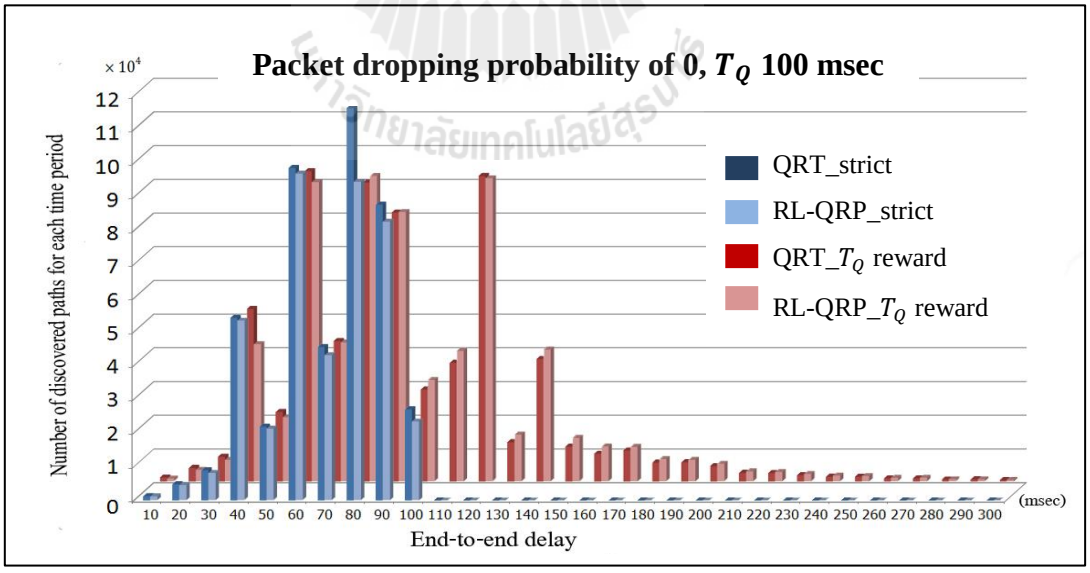

Figure 3.12 Number of discovered path under different end-to-end delays

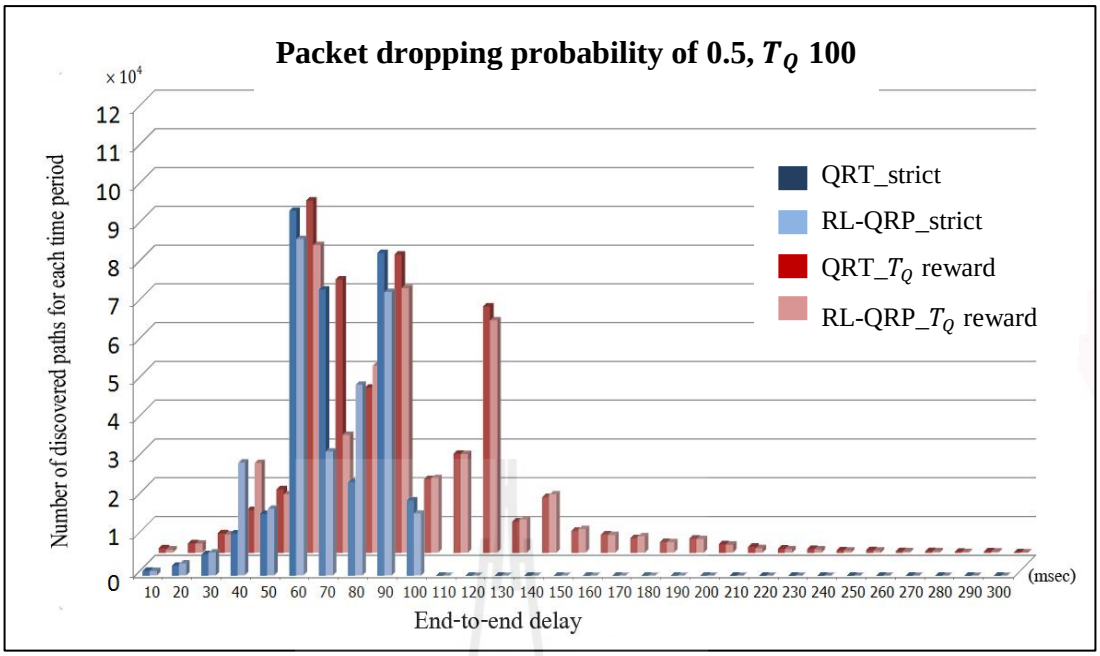

Figure 3.13 Number of discovered path under different end-to-end delays

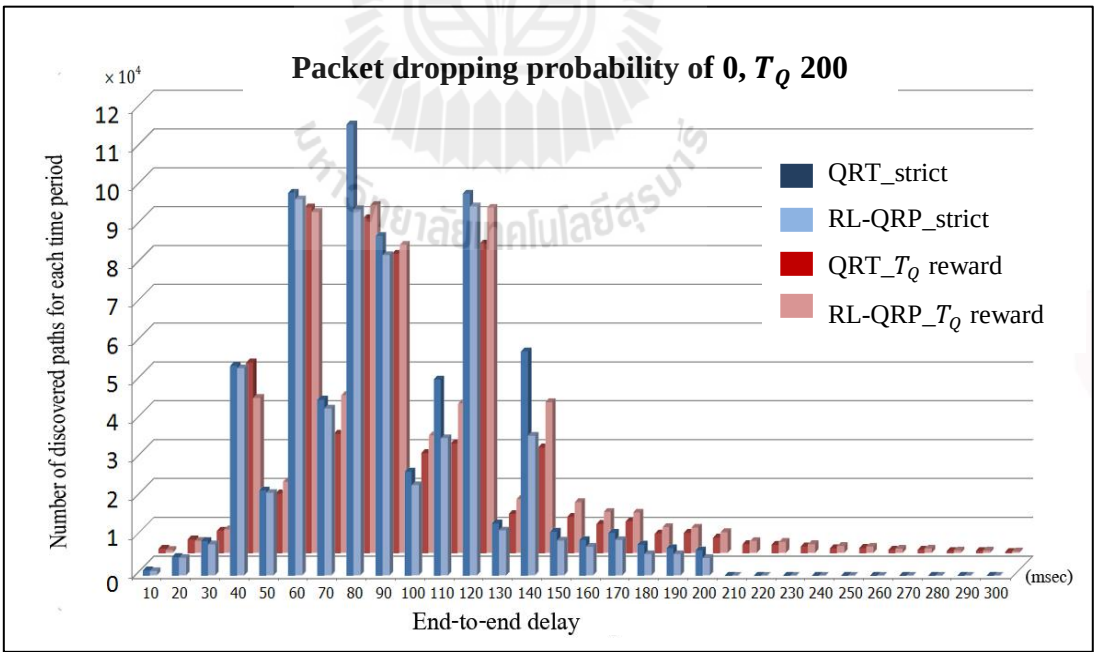

Figure 3.14 Number of discovered path under different end-to-end delays

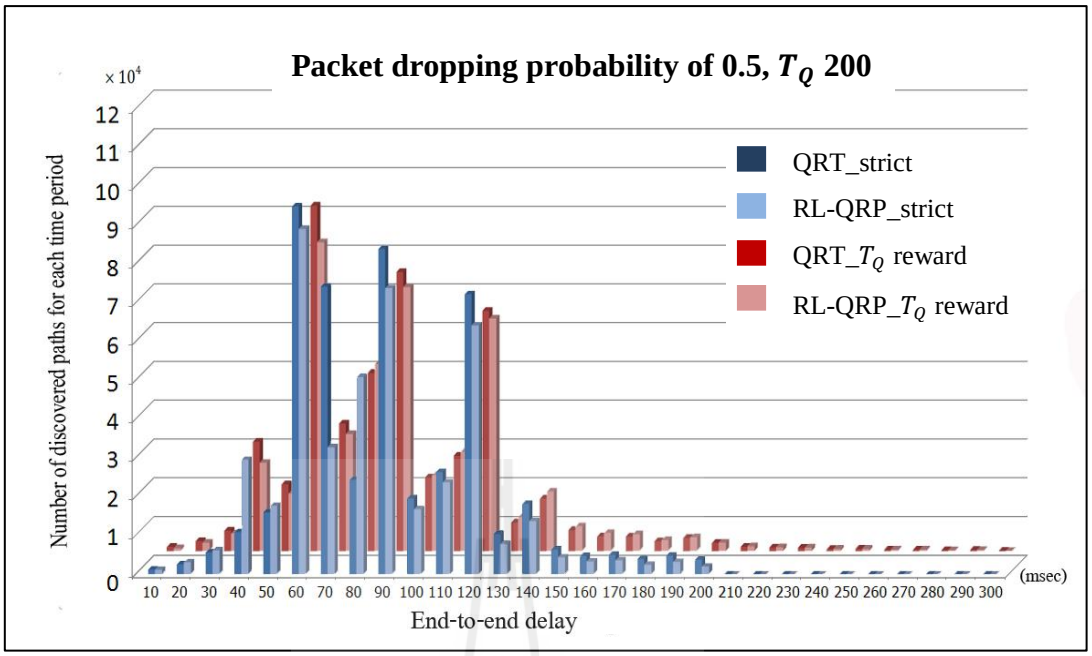

Figure 3.15 Number of discovered path under different end-to-end delays

3.6 Conclusion

We proposed the QRT routing algorithm for non-cooperative mWSNs which comprised of malicious stochastic packet dropping nodes. QRT was based on a RL routing method which incorporated a reputation and trust mechanism to screen out malicious nodes. The mechanism employed direct reputation from observed nodes to evaluate their trust values. We compared QRT against RL-QRP and threshold schemes. Results showed that the average success ratio of QRT was 11% and 25% greater than RL-QRP and the heuristic non-learning threshold schemes, respectively. As the mobility of the network increased, QRT consistently outperformed the other algorithms by gaining up to 9% and 22% success ratios above the RL-QRP and threshold schemes. The results suggest that reputation and trust mechanism can be applied to identify and avoid malicious packet dropping nodes mWSNs.

In terms of quality-of-service, the results have shown that QRT consistently outperformed RL-QRP even in presence of high packet dropping probability and stringent end-to-end delay requirements. The results suggest that QRT with reputation and trust mechanism scheme can be applied to cater quality-of-service in mWSNs.

CHAPTER IV

CONCLUSION AND FUTURE WORK

4.1 Conclusion

In this thesis, we proposed a routing method called QRT algorithm for non-cooperative mWSNs based on Reinforcement Learning (RL). In particular, the QRT was integration of a reputation and trust scheme to avoid misbehaving node with an existing RL-based routing protocol called RL-QRP. We evaluate its performance in non-cooperative mWSNs under various non-cooperation, mobility and end-to-end delay conditions. The experimental work carried out in this thesis was divided into two parts which were unconstrained and delay-constrained traffic demands. In the first experiment, we varied the number of malicious nodes and the number of mobile nodes to study their impact and compared the results with the original RL-QRP algorithm and a non-learning threshold scheme in terms of average success ratio (%), average end-to-end delay and the number of discovered path length. In the subsequent experiment, we then extended the framework to consider the delay-constrained quality-of-service into our simulation. We considered for 2 types of modification, including "QRT_strict" and "QRT_T_Q reward" and also compare the results with the same modifications on RL-QRP in terms of average success ratio (%), average end-to-end delay and the number of discovered paths under different end-to-end delay requirements. These two parts were presented in Chapter 3. The original contributions and findings in this thesis can be summarized as follows.

4.1.1 QRT

The first condition was the proposed QRT scheme which has that Q-learning algorithm can be applied to promote routing in mWSNs which include misbehaving nodes. We extended the state space which originally consisted of only the neighboring nodes of an agent to included quantized trust levels of their neighbors as well. We also modified the Q value updating equation (3.4) by adding $\frac{T_{ij}}{\alpha}$ as an additional reward term which reflected the trust between nodes, in order to take account of the trust level between nodes. Performance comparison was made with an existing RL-QRP algorithm and the threshold scheme. The simulation in the first part varied the number of malicious nodes along with the packet dropping probability of a malicious node. In the second part, the simulation varied the number of mobility nodes.

The proposed experiment results showed that the QRT method consistently outperformed RL-QRP and the threshold scheme in terms of success ratio when varying the number of malicious node and achieved up to 11% and 25%, respectively more than the two schemes. QRT method also discovered more longer paths than other schemes. When the number of mobility node increased, QRT gained up to 9% and 22% or success ratio over the RL-QRP algorithm and the threshold scheme, respectively.

4.1.2 Quality-of-Service

The purpose of this section was to add quality-of-service in terms of end-to-end delay requirement into our simulation. In the first part of this study, we modified the end-to-end delay requirement or T_Q value in the Q learning equation. Then, the results showed that varying T_Q alone cannot screen out path which had end-

to-end delay more than T_Q. An alternative approach was then trialed which selected next hop nodes whose path delay so far has not yet exceeded T_Q. The results suggested that QRT performed well in scenarios where end-to-end delay quality-of-service was required by the traffic demands even in the presence of malicious nodes, achieving up to 11% success ratio than RL-QRP.

The significance of our work was focused on proposing means to enhance routing in the presence of misbehaving nodes in mWSNs. We studied the effects of mobility and different degrees of malicious node behavior. Moreover, we added quality-of-service into the experiment for a more realistic biomedical application scenario using mWSNs. We can conclude that QRT approach can obtain the better routing performance than RL-QRP and the threshold scheme detecting and avoiding malicious nodes in mWSNs under various conditions of packet dropping probability, node mobility and stringent end-to-end delay requirements.

4.2 Future Work

4.2.1 mWSNs with Indirect Reputation Value

To study the effect of indirect reputation value which is the opinion about the next node by other neighbor nodes (for example, node *i* considers forwarding a packet to node *j*, then node *i* will get the opinion about node *j* by node k to evaluate trustworthiness of nodes *j*), Srinivasan, and Teitelbaum, (2006) proposed the distributed reputation-based beacon trust system (DRBTS) which used both direct reputation and indirect reputation based on beta distribution to weight reward for decision making of the node in choosing the next node. A possible direction for future extension of this thesis is therefore to include indirect reputation in the framework.

4.2.2 Traffic Priority

In biomedical mobile wireless sensor networks, there is a great variety of health information and the significance of each information is different. Giving priority to important information such as heart rate over delay tolerable traffic such as temperature 200-300 msec by reserving short routes for only important information to avoid packet collision and over buffering. Hence, traffic and route prioritization are promising directions for further study.

4.2.3 Performance Evaluation of Test Bed

The main objective of this thesis was to improve routing performance in mWSNs by using RL with trust and reputation. This experiment was simulated in Visual C++ environment to perform the learning process and evaluate algorithms. Therefore, an important future direction is to extend towards real data collection for training the learning algorithm in actual mWSNs.

4.2.4 mWSNs with Energy Consumption Condition

Energy consumption in mWSNs is one of the most important issues. Dealing with the energy problems in mWSNs by expanding the state space of remaining battery of each node and making energy-aware routing decisions at intermediate nodes along the route warrants further investigation.

REFERENCES

Aghaei, R., Rahman, A., Gueaieb, W., Saddik, A. (2007). Ant Colony-Based Reinforcement Learning Algorithm for Routing in Wireless Sensor Networks. **Proceedings of Instrumentation and Measurement Technology Conference**.

Blaze, M., Feigenbaum, J., Lacy, J. (1996). Decentralized Trust Management. **Proceedings of Security and Privacy**.

Buchegger, S., Boudec. J.Y. (2002). Performance Analysis of the CONFIDANT Protocol (Cooperation of Nodes-Fairness in Dynamic Ad-hoc NeTworks), **Proceedings of The Third ACM International Symposium on Mobile Ad Hoc Networking and Computing**.

Buchegger, S., Boudec. J.Y.L. (2003a). Coping with False Accusations in Misbehavior Reputation Systems for Mobile in Ad-Hoc Networks. **Technical Report IC, 2003, 31, EPFL-DI-ICA.**

Buchegger, S., Boudec. J.Y.L. (2003b). The Effect of Rumor Spreading in Reputation Systems for Mobile Ad-hoc Networks. **Proceedings of Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks**.

Chen, H., Wu, H., Zhou, X., Gao, C. (2007). Reputation-based Trust in Wireless Sensor Networks, **Proceedings of Multimedia and Ubiquitous Engineering**.

Dong, S., Agrawal, P., Sivalingam, K. (2007). Reinforcement Learning Based Geographic Routing Protocol for UWB Wireless Sensor Network. **Proceedings of Global Telecommunications Conference**.

Forster, A., Murphy, A.L. (2007). Exploiting Reinforcement Learning for Multiple Sink Routing in WSNs. **Proceedings of National Competence Center in Research on Mobile Information and Communication Systems**.

Forster, A., Murphy, A.L., Schiller, J., Terfloth, K. (2008). An Efficient Implementation of Reinforcement Learning Based Routing on Real WSN Hardware, **Proceedings of International Conference on Wireless and Mobile Computing**.

Ganeriwal, S., Srivastava, M. B. (2004). Reputation based Framework for High Integrity Sensor Networks, **Proceedings of Security of Ad Hoc and Sensor Networks**.

Gelenbe, E., Gellman, M. (2007). Oscillations in a Bio-Inspired Routing Algorithm, **Proceedings of Mobile Ad Hoc and Sensor Systems**.

He, D., Chen, C., Chan, S., Bu, J., Vasilakos, A. (2012). ReTrust: Attack-resistant and Lightweight Trust Management for Medical Sensor Networks, **Journal of Information Technology in Biomedicine**, vol. 16, no. 4, pp. 623-632.

Istepanian, R.S.H., Jovanov, E., Zhang, Y.T. (2004). Guest Editorial Introduction to the Special Section on M-Health: Beyond Seamless Mobility and Global Wireless Health-Care Connectivity, **Journal of Information Technology in Biomedicine**, vol. 8, no.4, pp. 405-414.

Josang, A., Knapskog, S.J. (1998). A Metric for Trust Systems, **Proceedings of The 21st National Information Systems Security Conference**.

Jovanov, E., Poon, C., Guang-Zhong, Y., Zhang, Y.T. (2009). Guest Editorial Body Sensor Networks: From Theory to Emerging Applications. **Journal of Information Technology in Biomedicine,** vol. 13, no. 6, pp. 859-863.

Kaelbling, L.P., Littman, M.L., Moore, A.P. (1996). Reinforcement Learning: A survey. **Journal of Artificial Intelligence Research**, vol. 4, pp. 237-285.

Karaki, J.N., Kamal, A.E., (2004). Routing Techniques in Wireless Sensor Networks: a Survey. **Journal of Wireless Communications**, vol. 11, no. 4, pp. 6-28.

Kim, K., Kim, H. Hong, Y. (2009). A Self Locolization Scheme for Mobile Wireless Sensor Networks**., Proceedings of Computer Sciences and Convergence Information Technology**.

Kim, K., Lee, I.S., Yoon, M., Kim, J., Lee, H., Han, K. (2009). An Efficient Routing Protocol Based on Position Information in Mobile Wireless Area Body Sensor Network. **Proceedings of Networks and Communications**.

Lan Tien Nguyen, Defago, X., Beuran, R., Shinoda, Y. (2008). An Energy Efficient Routing Scheme for Mobile Wireless Sensor Networks. **Proceedings of Wireless Communication Systems**.

Michiardi, P., Molva, R. (2002). Core: A Collaborative Reputation Mechanism to Enforce Node Cooperation in Mobile Ad hoc Networks. **Proceedings of Communications and multimedia Security**.

Pang, Z., Chen, Q., Zheng, L. (2009). A Pervasive and Preventive Healthcare Solution for Medication Noncompliance and Daily Monitoring. **Proceedings of Applied Sciences in Biomedical and Communication Technologies**.

Puterman, M. (1994). **Markov Decision Processes: Discrete Stochastic Dynamic Programming**: Wiley-Interscience.

Renaud, J.C., Tham, C.K. (2006). Coordinated Sensing Coverage in Sensor Networks using Distributed Reinforcement Learning**., Proceedings of International Conference on Networks**.

Resnick, P., Zeckhauser, R. (2000). Trust among strangers in Internet transactions: Empirical analysis of eBays Reputation System. **Journal of The Economics of the Internet and E-commerce: Advance in applied microeconomics**, vol. 11, pp. 127-157.

Resnick, P., Kuwabara, K., Zeckhauser, R., Friedman, E. (2000). Reputation systems. **The Article of Communications of the ACM**, vol. 43, no.12, pp. 45-48.

Seah, M.W.M., Tham, C.K., Srinivasan, V., Xin, A. (2007). Achieve Coverage through Distributed Reinforcement Learning in Wireless Sensor Networks. **Proceedings of Intelligent Sensor, Sensor Networks and Information**.

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

Tanachaiwiwat, S., Dave, P., Bhindwale, R., Helmy, A. (2003). Location-centric Isolation of Misbehavior and Trust Routing in Energy-Constrained Sensor Networks. **Proceedings of Performance, Computing and Communications**.

Vashney, U. (2008). Improving Wireless Health Monitoring Using Incentive-Based Router Cooperation, **IEEE Computer Magazine**, vol. 41, no. 5, pp. 56-62.

Wang, P., Wang, T. (2006). Adaptive Routing for Sensor Networks using Reinforcement Learning. **Proceedings of Computer and Information Technology**.

Watkins, C. 1989. **Learning from Delayed Rewards**. University of Cambridge, England.

Xuedong, L., Balasingham, I., Byun, S.S. (2008). A Multi-agent Reinforcement Learning based Routing Protocol for Wireless Sensor Networks. **Proceedings of Wireless Communication Systems**.

Xuedong, L., Balasingham, I., Byun, S.S. (2008). A Reinforcement Learning based Routing Protocol with QoS Support for Biomedical Sensor Networks. **Proceedings of Applied Sciences on Biomedical and Communication Technology**.

Yadav, V., Mishra, M.K., Gore, M.M. (2009). Localization Scheme for Three Dimensional Wireless Sensor Networks Using GPS enabled Mobile Sensor Nodes. **Journal of Next-Generation Networks**, vol. 1, no. 1, pp. 60-72.

Ying-Hong, W., Chin-Yung, Y., Wei-Ting, C., Chun-Xuan W. (2008). An Average Energy based Routing Protocol for Mobile Sink in wireless sensor networks. **Proceedings of Ubi-Media Computing**.

Yu, H., Shen, Z., Miao, C., Leung, C., Niyato, D. (2010). A Survey of Trust and Reputation Management Systems in Wireless Communications. **Journal of the IEEE**, vol. 98, no. 10, pp. 1755-1772.

Zhou, Y., Xing, J., Yu, Q. (2006). Overview of Power-efficient MAC and Routing Protocols for Wireless Sensor Networks. **Proceedings of Mechatronic and Embedded Systems and Applications**.

APPENDIX

PUBLICATION

Publication

Naputta, Y., and Usaha, W. (2012). **RL-based Routing in Biomedical Mobile Wireless Sensor Networks using Trust and Reputation.** The 9th International Symposium on Wireless Communication Systems (ISWCS), France, August 2012.

RL-based Routing in Biomedical Mobile Wireless Sensor Networks using Trust and Reputation

Yanee Naputta
School of Telecommunication Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand 30000
E-mail: yanee_naputta@hotmail.com

Wipawee Usaha
School of Telecommunication Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand 30000
E-mail: wusaha@ieee.org

Abstract— **The main function of biomedical sensor network is to guarantee that the data packets from patients can be delivered reliably to the destination node or medical center. Attached to patients, these nodes can be mobile, thus forming a mobile wireless sensor network (mWSN). Moreover, non-cooperative nodes may also be present in the network. This paper therefore proposes a routing method for non-cooperative mWSNs based on Reinforcement Learning (RL). In particular, a reputation and trust scheme to avoid misbehaving nodes was integrated with an existing RL-based routing protocol called RL-QRP. We evaluated its performance in non-cooperative mWSNs under various conditions of non-cooperation and mobility. We found that the proposed method can achieve a success ratio of up to 11% over the RL-QRP, and 25% over a non-learning brute force search threshold method.**

Keywords- Reinforcement Learning; Mobile Wireless Sensor Networks; Routing;Non-cooperative; Trust and reputation

I. INTRODUCTION

In this paper, routing issues in biomedical wireless sensor networks are investigated. Parameters such as body temperature, blood pressure heart rate are sensed at a patient and transmitted via intermediate sensor nodes to a base station at a medical center. The data is used for health status monitoring, diagnosis and treatment. For example [1], [2] proposed the use of wireless sensors to monitor vital signs of patients in hospital and home environments.

The most important task of biomedical sensor networks is to ensure that data can be delivered to the medical center reliably and efficiently [3]. Furthermore, in biomedical sensor networks, patients may be moved to an emergency room, and medical staff may want to know their information continuously. Therefore, use of a mobile wireless sensor network (mWSN) is necessary for biomedical sensors networks. A distributed, lightweight, and highly adaptive routing protocol based on methods such as reinforcement learning (RL) has been proposed for such rapidly changing wireless network conditions [4], [5].

RL is a technique that has been used to support routing in dynamic topology networks. RL is a study of how artificial systems can learn to optimize their behavior by using its experience through rewards and punishments. There are some works which applied RL to solve routing problem in static WSNs [6]. In [4], the authors proposed a Cognitive Packet Network (CPN) which made routing decisions in presence of routing oscillations using RL and a neural network model. Ref. [5] proposed RL-QRP, a RL-based routing protocol with routing scheme in mWSNs. They investigated the impact of network traffic load and sensor node mobility on the network performance. However, their results were based on the assumption that all nodes cooperated in the packet forwarding process. But a more realistic scenario would require consideration of situation which some nodes do not cooperate with each other (i.e., by dropping packets they receive) either due to node battery depletion, malfunctioning or simply misbehaving for unknown reason [7]. Since in biomedical sensor networks, data packets must be delivered to its destination node reliably, means to identify and avoid these malicious nodes are necessary [8].

Reputation and trust schemes have been used to identify well-behaved and malicious nodes in WSNs [8], [9]. In such schemes, a sensor node continuously builds a reputation value for other nodes by monitoring their behavior. Then the sensor node uses this reputation value to evaluate the trustworthiness of other nodes. Ref. [8] proposed a trust scheme called ReTrust for medical WSNs which is lightweight and attack-resistant. High malicious node detection rates and average packet delivery ratio were achieved via simulation and experimental test-bed. However, sensor node mobility was not explicitly addressed.

Therefore, the objective of this paper is to solve the routing problem for non-cooperative mWSNs based on RL by incorporating a reputation and trust mechanism which screens out nodes with malicious behavior using values of reputation and trust values maintained at the sensor nodes. We compared its performance with an existing reinforcement learning routing scheme called RL-QRP [5] under various mobility and malicious node scenarios.

978-1-4673-0762-8/12/$31.00 ©2012 IEEE

II. RELATED WORKS

A. RL-QRP

Reinforcement Learning based Routing Protocol with QoS Support for Biomedical Sensor Networks (RL-QRP) has been proposed for promote routing policies to find optimal path through experience and rewards [5]. They used Q-learning which learns the value of function $Q(s,a)$ to find an optimal decision policy. In each time action a is selected, the agent receives an immediate reward r from the environment. Then the agent will use this reward to update the one step rule as follows

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a')] \quad (1)$$

where the Q-value, $Q(s,a)$, denotes the quality of action a at state s, α is the learning rate and γ is the discount factor. $Q(s',a')$ denotes the expectation future reward at state s' by taking action a'. The updated Q-values then in turn affect the future decisions of the agent.

RL-QRP requires the use of location information parameters to calculate a reward following a particular action. Therefore, the protocol can find the shortest path from a beginning node to a destination node using a reward function given by

$$r = \begin{cases} \left(\frac{(Ds_i,s_{sink} - Ds_j,s_{sink})}{Ds_i,s_{sink}}\right) / \left(\frac{Ts_i,s_j}{T_Q}\right), & ACK\ received \\ -\frac{Ds_i,s_j}{Ds_i,s_{sink}}, & ACK\ not\ received, \end{cases} \quad (2)$$

where Ds_i,s_{sink} and Ds_j,s_{sink} is the distance between node s_i, s_j and destination node s_{sink}, respectively. Ds_i,s_j is the distance between node s_i and node s_j. T_Q is the end-to-end delay requirement encapsulated in the data packet. Ts_i,s_j is the experience delay between node s_i and s_j.

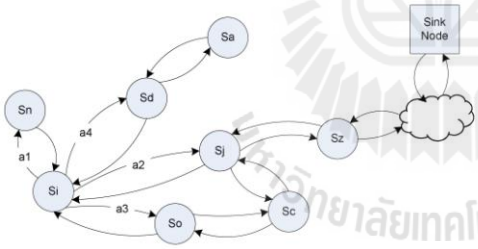

Figure 1. RL-QRP routing model

The basic idea of RL-QRP follows Figure 1. Each node in the biomedical sensor network is considered as a state belonging to set S = $\{s_i\}$, i = 1,2,…,N where N is the number of sensor nodes. For each node with a neighbor s', an action can be selected from A = $\{a(s_j|s_i)\}$. Note that $a(s_j|s_i)$ refers to a packet being forwarded from state s_i to s_j, provided that s_i and s_j are within each other's communication range. Suppose that node s_i in Figure 1 must forward a packet to the sink node through some intermediate node. s_i then checks the Q-value of its neighboring nodes which include s_n, s_j, s_o, s_d. Then node s_i forwards the packet to the neighbor node with the highest Q-value. Suppose that s_i forwards the packet to node s_j. After that node s_i updates its Q-value $Q\left(s_i, a(s_j|s_i)\right)$ according to (1) with reward in (2). The process is repeated for node s_j and the following consecutive nodes until the packet reaches the sink node. Thus, the nodes can find the optimal route through experience and rewards without complicated prediction techniques, or explicitly frequently updating. Therefore, this process is well-suited for dynamic topologies.

B. Reputation

Reputation and trust systems have been proved useful mechanisms to address the threat of compromised or faulty entities. Such systems are operated by identifying selfish peers and excluding these entities from the networks. Ref. [10] considered routing protocols in MANETS by using both first hand and second hand information for updating reputation values. Ref. [11] and [8] considered both first and second hand reputation and trust-based models developed exclusively for sensor networks. In [8], a two-tier architecture trust management scheme was proposed in which a master node was used to compute the trust values for sensor nodes within its range. In [11], a watchdog mechanism was used to build their trust rating system. Given a reputation value R_{ij} obtained from the watchdog, the trust metric T_{ij} based on BETA distribution [9] can be computed by

$$T_{ij} = E[R_{ij}] = \frac{p+1}{p+n+2} \quad (3)$$

where T_{ij} refers to node s_i's prediction of the expected future behavior of positive outcomes of node s_j, p and n are the number of positive and negative outcomes of a specific event, respectively. In particular,p and n are the number of successes and failures in forwarding packets between two nodes, respectively. The first hand or direct reputation value can be determined from $< p, n >$ which is the direct observation of node s_j (the observed node) experienced by node s_i. From Figure 1, suppose that node s_i prefers to forward the data packet to the destination node by the shortest path via node s_j and s_z. In effect, an interaction occurs between node s_i and node s_j. We used a simply binary rating scheme, where a successful outcome (p) is incremented if node s_j forwards the packet to node s_z and a failed outcome (n) is incremented if node s_j does not forward the packet to node s_z. Note that typically $p, n \geq 0$ so that the trust value is normalized to the range $[0,1]$, and the initial value of trust is 0.5. On the other hand, the indirect reputation value can be determined from direct reputation values of node s_j recommended by its neighboring nodes. Although aggregated second hand information (i.e. by inquiring from watchdog the values of $< p, n >$ of other nodes which interacted with node s_j in the past) helps accelerate the calculation of the reputation value, this paper considers the first hand observation or direct reputation for the sake of simplicity. Furthermore, drawbacks of indirect reputation include vulnerability to bad-mouthing attacks and that watchdog may not be able to capture all relevant information in the network [9].

III. RL-QRP WITH REPUTATION AND TRUST

In this section, the proposed RL-based routing integrated with reputation and trust, called QRT, is described. We redefine the state and action and rewards as follows:

a) Let Q(s_i, s_j, l_{ij}) denote the opinion of s_i about s_j which is updated when node s_j forwards or drops packets to its neighboring node:

$$Q(s_i, s_j, l_{ij}) = 0.5\left\{(1-\alpha)Q(s_i, s_j, l_{ij}) + \alpha\left[r + \frac{T_{ij}}{\alpha} + \left(\gamma \max_{s'_j} Q(s'_i, s'_j, l'_{ij})\right)\right]\right\}, \quad (4)$$

where the Q-value, $Q(s_i, s_j, l_{ij})$, denotes the quality of forwarding packets at node s_i experienced by s_i and l_{ij} denotes the level of trust at node s_i experienced by $s_i$which is quantized into 10 levels (level 0-9). A trust value T_{ij}which takes values in the range [0,1] is quantized uniformly into intervals of 0.1.

b) State: S = $\{s_i\}$, i = 1,2,…,N where N is the number of sensor nodes. Each node is a state in S.

c) Trust: T_{ij} is the trust value that quantifies the trustworthiness of s_j in forwarding packets from node s_i.

d) Action: $A = \{a(s_j|s_i)\}$, s_i , s_j $\in S$. Excution of $a(s_j|s_i)$ means that the packet is forwarded from state s_i to s_j, provided that $s_i$and s_j are within each other's communication range.

e) Reward function: r is the reward for executing an action at node s_i (e.g. s_i forwards the packet to s_j) given by:

$$r = \left(\frac{(Ds_i,s_{sink} - Ds_j,s_{sink})}{Ds_i,s_{sink}}\right) / \left(\frac{Ts_i,s_j}{T_Q}\right). \quad (5)$$

Note that we assumed that every node in the network always sends ACK back to its upstream node, regardless of their behavior. Ds_i, s_{sink} and Ds_j, s_{sink} are the distance between node s_i , s_j and the destination node s_{sink} , respectively. Ds_i, s_j is the distance between node s_i and node s_j. T_Q is the end-to-end delay requirement encapsuled in the data packet. Ts_i, s_j is the experienced delay between node s_i and s_j.

The pseudo code of the proposed QRT routing algorithm is shown in Table I.

IV. PERFORMANCE AND EVALUATION

In this section, we evaluated the proposed QRT routing algorithm which integrated the existing RL-QRP [5] with the reputation and trust scheme. Results were compared with the original RL-QRP and a non-learning threshold reputation scheme. The latter scheme ranked the trust values of the neighboring nodes and selected the next node with the highest trust value above a predetermined threshold of 0.4 which was found to give the best performance among other threshold values. Visual C++ was used to simulate a mWSN under various conditions according to Table II. A number of nodes in the mWSN were mobile and followed the random way point mobility model with velocity randomly chosen from [0,5] m/s. The remaining nodes were assumed static. Each experiment was repeatedly run with different seeds, each with a runlength of 10^6 events until the sample averaged results were within a 10% range. The remaining simulation parameters are shown in Table II.

TABLE I. PSEUDO CODE

```
01 Begin
02 Initialization
03    Set timer for beacon exchange
04 Begin Loop
05 If timer expires
06    Broadcast beacon to immediate neighboring nodes
07    Re-set timer
08 Endif
09 If beacon packets arrives
10    Update neighboring node's position and Q-value
11 Endif
12 If data packet arrives
13       If good node
14          Random number
15             If Random number > ε
16                Select neighboring node with highest Q-value
17             Else
18                Randomly select neighboring node
19             End if
20          Receive reward r
21          Update the Q-value
22          Update Trust
23       Else
24          Drop packets
25       End if
26 Endif
27 Go to 04
28 End
```

TABLE II. SIMULATION PARAMETERS

Parameters	Value	
	Part 1	Part 2
Number of sensor nodes	36	
Node mobility	Random way point	
Pause time (s)	60	
Node velocity (m/s)	Min. 0, Max. 5	
Area size	200x200m^2	
Transmission range	50m	
Runlength (number of route requests)	10^6	
Learning rate (α) for RL-QRP, QRT	0.5	
Discount factor (γ) for RL-QRP, QRT	0.5	
Number of mobile nodes	9	0,9,18,27,36
Number of malicious nodes	9, 18	9
Probability of dropping a packet	0, 0.25,0.5,0.75, 1	0.25

A. Part 1 Malicious Nodes Effect

In this experiment, there are 9 mobile nodes out of 36 nodes. To study the effect of malicious nodes and the degree to which they misbehave, the number of malicious node was varied from 9 to 18 nodes and their packet dropping probability were varied from 0 to 1. The following metrics were measured:

- **Average success ratio (%)** is given by:

$$Average\ success\ ratio = \frac{number\ of\ discovered\ paths}{number\ of\ routing\ requests} x100 \quad (7)$$

This metric is the proportion of successfully discovered paths. Figure 2 illustrates the average success ratio for QRT, RL-QRP and threshold schemes as the packet dropping probability was varied. Note that for all packet dropping probabilities, the average success ratio of QRT was up to 11% greater than RL-QRP and up to 25% greater than the threshold scheme. Such result indicated that QRT can identify and avoid malicious nodes more effectively than RL-QRP and threshold schemes and thereby discover more paths that can reach the destination node.

- **Average end-to-end delay:** In Figure 3, the average end-to-end delay is shown against the packet dropping probability. Note that the QRT showed a higher average end-to-end delay than RL-QRP. This was because QRT can discover more paths than the other schemes as shown in the previous figure. In Figure 4, such paths included both short paths (2, 3 hops) which were comparable to the RL-QRP, as well as long paths (4 hops up) which were discovered significantly greater than RL-QRP. The threshold scheme discovered the least number of shortest paths of all thus obtaining the highest average end-to-end delay.

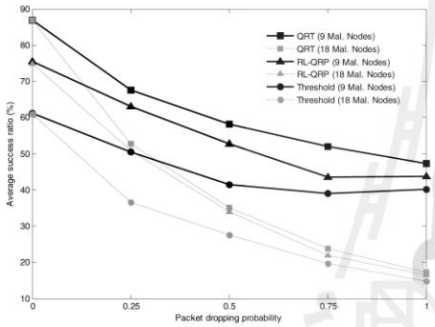

Figure 2. Average success ratio of discovered paths

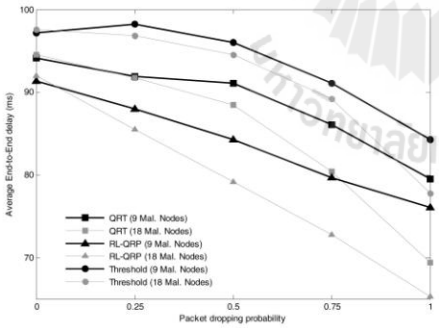

Figure 3. Average end-to-end delay of discovered paths

B. Part 2 Mobility Effect

In this part, the algorithms' performance when varying node mobility was investigated. For this scenario, 9 malicious nodes were present, each with a packet dropping probability of 0.25. Such setting was used because high success ratio were observed for all schemes. Hence, the effect from increased mobility would be more visible. The degree of mobility was varied by increasing the number of moving nodes from 0 (the least mobile) to 36 (the most mobile).

- **Average success ratio (%):** Figure 5 illustrates the average success ratio for all schemes. Note that QRT consistently outperformed both RL-QRP and threshold schemes by up to 9% and 22%, respectively. However, the margin between QRT and RL-QRP decreased as mobility increased.

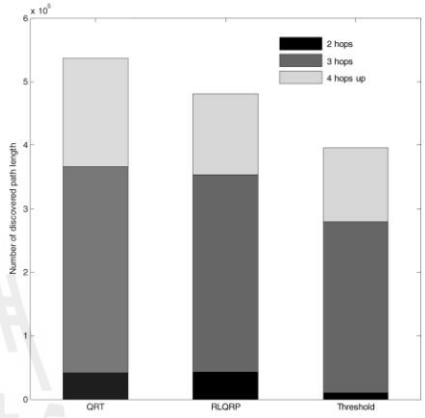

Figure 4. Number of discovered paths for each path length for 9 malicious nodes

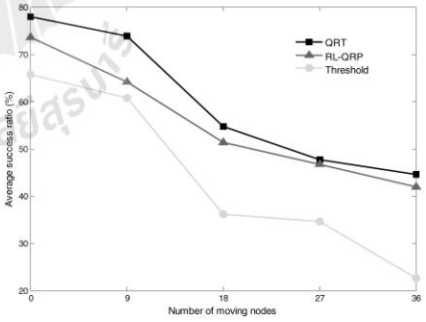

Figure 5. Average success ratio under various degrees of mobility

- **Average end-to-end delay:** In Figure 6, the average end-to-end delay is shown versus the number of moving nodes. Similar to Figure 3, the average end-to-end delay of QRT was greater than RL-QRP but less than the threshold scheme. This was because, in Figure 7, QRT can find more longer paths (4 hops up) than RL-QRP and the threshold scheme, while obtaining a comparable number of short paths (2, 3 hops) to RL-QRP. Furthermore, as the number of discovered paths gradually decreased as mobility increased, QRT consistently discovered more paths than other schemes.

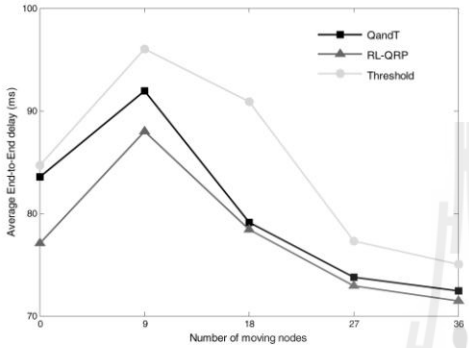

Figure 6. Average end to end delay of discovered paths under various degrees of mobility

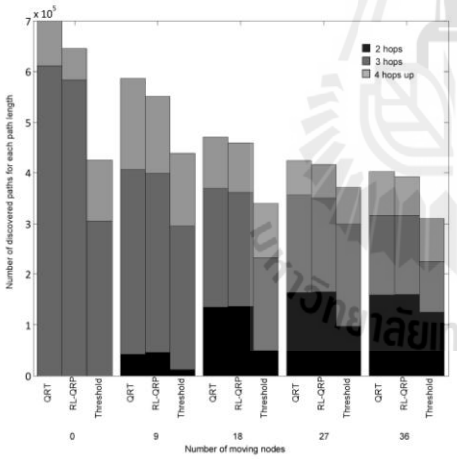

Figure 7. Average number of discovered path length under various degrees of mobility

V. Conclusion

We proposed the QRT routing algorithm for non-cooperative mWSNs which comprised of malicious stochastic packet dropping nodes. QRT was based on a RL routing method which incorporated a reputation and trust mechanism to screen out malicious nodes. The mechanism employed direct reputation from observed nodes to evaluate their trust values. We compared QRT against RL-QRP and threshold schemes. Results showed that the average success ratio of QRT was up to 11% and 25% greater than RL-QRP and the heuristic non-learning threshold schemes, respectively. As the mobility of the network increased, QRT consistently outperformed the other algorithms by gaining up to 9% and 22% success ratios above the RL-QRP and threshold schemes. The results suggest that reputation and trust mechanism can be applied to identify and avoid malicious stochastic packet dropping nodes mWSNs.

Our future work will consider QoS issues such as delay requirements, as well as the impact of interference and QRT's communication and computation overhead, and performance comparison with indirect reputation values.

References

[1] Z. Pang, Q. Chen, L. Zheng, "A Pervasive and Preventive Healthcare Solution for Medication Noncompliance and Daily Monitoring", International Symposium on Applied Sciences in Biomedical and Communication Technologies, pp. 315-320, 2009.

[2] E. Jovanov, C. Poon, Y. Guang-Zhong, Y.T. Zhang, "Guest Editorial Body Sensor Networks: From Theory to Emerging Applications", IEEE Transactions on Information Technology in Biomedicine, Vol. 13, No. 6, pp. 859-863, 2009.

[3] R.S.H. Istepanian, E. Jovanov, Y.T. Zhang, "Guest Editorial Introduction to the Special Section on M-Health: Beyond Seamless Mobility and Global Wireless Health-Care Connectivity", IEEE Transactions on Information Technology in Biomedicine, Vol. 8, No. 4, pp. 405-414, 2004.

[4] E. Gelenbe, M. Gellman, "Oscillations in a Bio-Inspired Routing Algorithm", IEEE International Conference on Mobile Ad Hoc and Sensor Systems, pp. 1-7, 2007.

[5] L. Xuedong, I. Balasingham, S.S. Byun, "A Reinforcement Learning based Routing Protocol with QoS Support for Biomedical Sensor Networks", International Symposium on Applied Sciences on Biomedical and Communication Technology, pp. 1-5, 2008.

[6] A. Forster, A.L. Murphy, J. Schiller, K. Terfloth, "An Efficient Implementation of Reinforcement Learning Based Routing on Real WSN Hardware", IEEE International Conference on Wireless and Mobile Computing, pp. 247-252, 2008.

[7] U. Vashney, "Improving Wireless Health Monitoring Using Incentive-Based Router Cooperation", IEEE Computer Magazine, Vol. 41, No. 5, pp. 56-62, 2008.

[8] D. He, C. Chen, S. Chan, J. Bu, A. Vasilakos, "ReTrust: Attack-resistant and Lightweight Trust Management for Medical Sensor Networks", IEEE Transactions on Information Technology in Biomedicine, Online, May 2012.

[9] H. Yu, Z. Shen, C. Miao, C. Leung, D. Niyato, "A Survey of Trust and Reputation Management Systems in Wireless Communications", Proceedings of the IEEE, Vol. 98, No. 10, pp. 1755-1772, 2010.

[10] S. Buchegger, J.-Y. Le Boudec., "Performance Analysis of the CONFIDANT Protocol (Cooperation of Nodes-Fairness in Dynamic Ad-hoc NeTworks)", International Symposium on Mobile Ad Hoc Networking and Computing, 2002.

[11] S. Ganeriwal, M. B. Srivastava, "Reputation based Framework for High Integrity Sensor Networks", ACM Workshop on Security of Ad Hoc and Sensor Network, pp. 66-77, 2004.

BIOGRAPHY

Ms. Yanee Naputa was born on July 15, 1986 in Nakhon Ratchasima province, Thailand. She finished high school education from Boonwattana School, Nakhon Ratchasima province. She received her Bachelor's Degree in Engineering (Telecommunication) from Suranaree University of Technology in 2009. For her post-graduate, she continued to study for her Master's degree in the School of Telecommunication Engineering, Institute of Engineering, Suranaree University of Technology. During her Master's degree education, she was a visiting researcher at the Centre for Dynamic Intelligent Communication (CIDCOM), Department of Electrical and Electronic Engineering, University of Strathclyde, Scotland.