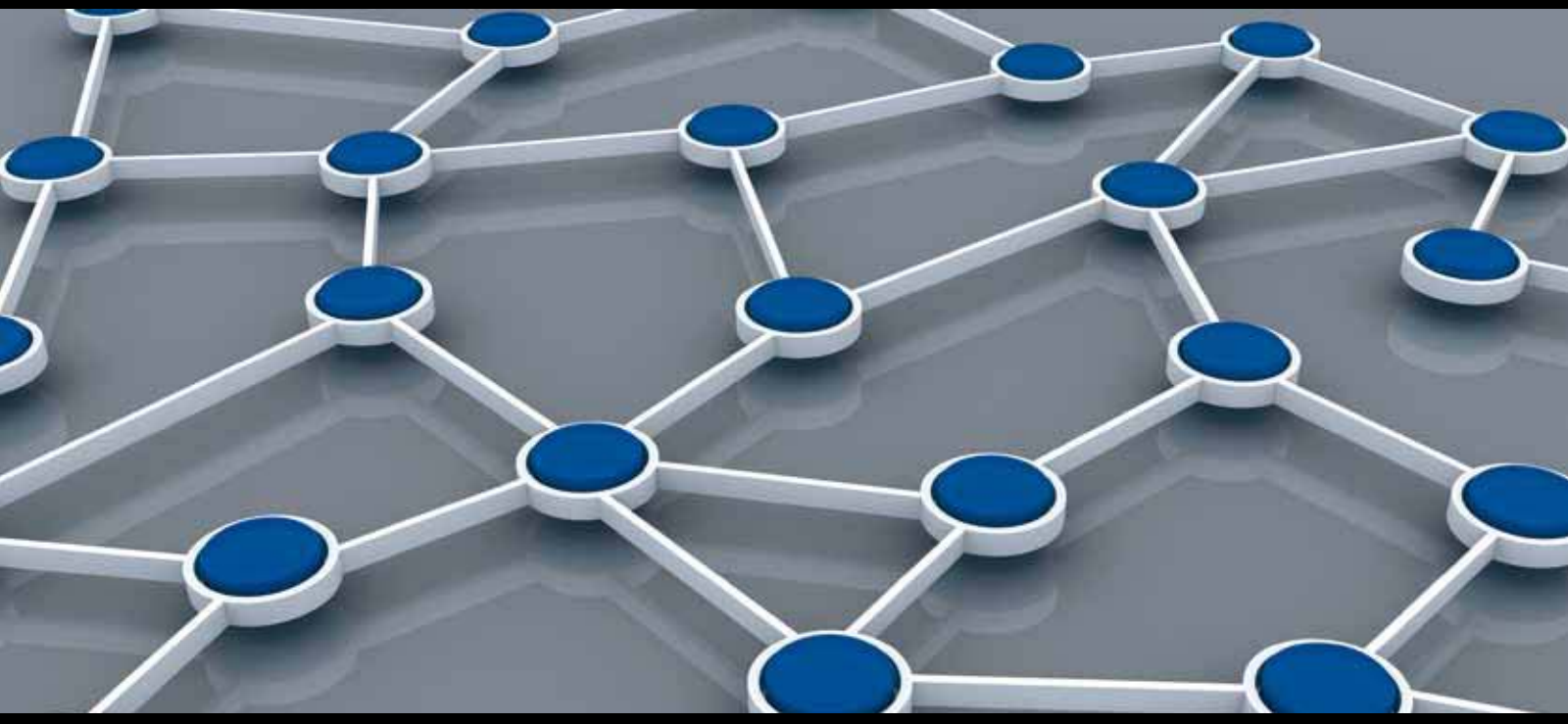
International Journal of Distributed Sensor Networks

Sensor Networks for High-Confidence Cyber-Physical Systems

Guest Editors: Feng Xia, Tridib Mukherjee, Yan Zhang, and Ye-Qiong Song



Sensor Networks for High-Confidence Cyber-Physical Systems

International Journal of Distributed Sensor Networks

Sensor Networks for High-Confidence Cyber-Physical Systems

Guest Editors: Feng Xia, Tridib Mukherjee, Yan Zhang, and Ye-Qiong Song

Editorial Board

Contents

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 245734, 3 pages
doi:10.1155/2011/245734

Editorial
Sensor Networks for High-Confidence Cyber-Physical Systems

Feng Xia,[1] Tridib Mukherjee,[2] Yan Zhang,[3] and Ye-Qiong Song[4]

[1] *School of Software, Dalian University of Technology, Dailan 116620, China*
[2] *Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85281, USA*
[3] *Simula Research Laboratory and University of Oslo, 0316 Oslo, Norway*
[4] *Nancy University and LORIA, 54516 Vandoeuvre-lès-Nancy, France*

Correspondence should be addressed to Feng Xia, f.xia@ieee.org

Received 21 November 2011; Accepted 21 November 2011

Technical advances in ubiquitous sensing, embedded computing, and wireless communication are leading to a new generation of engineered systems called cyber-physical systems (CPSs). This field is attracting more and more attention from researchers and practitioners, as well as the governments. Technically, cyber-physical systems are integrations of computation, networking, and physical dynamics, in which embedded devices are massively networked to sense, monitor, and control the physical world. CPS has been regarded as the next computing revolution. This revolution will be featured by the envisioned transform that CPS will make on how we interact with the physical world.

To facilitate unprecedented interactions between human beings and the physical world, sensor networks will become a crucial ingredient of CPS due to the need for coupling geographically distributed computing devices with physical elements. The proliferation of affordable sensor network technologies has significantly contributed to recent progress in CPS. On the other hand, the unique features of CPS (e.g., cyber-physical coupling driven by new demands and applications) give rise to a lot of open challenges for design and deployment of sensor networks in such systems. In particular, high-confidence CPS requires the employed sensor networks to support real-time, dependable, safe, secure, and efficient operations.

In this issue we are focused on latest research results in wireless sensor networks (WSNs) that address key issues related to high-confidence cyber-physical systems and applications. In response to our call for papers, we have received 36 submissions, out of which 15 papers are finally accepted as a result of thorough review process by international experts in respective areas. The selection provides a glimpse of the state-of-the-art research in the field.

In the paper "*A game theoretic approach for interuser interference reduction in body sensor networks*," G. Wu et al. present a decentralized interuser interference reduction scheme with noncooperative game for body sensor networks (BSNs). A no-regret learning algorithm for reducing the effect of the interuser interference with low power consumption is proposed. The correctness and effectiveness of the proposed scheme are theoretically proved, and experimental results demonstrate that the effect of interuser interference can be reduced effectively with low power consumption.

The paper "*Low-complexity, distributed characterization of interferers in wireless networks*" by V. Kapnadak et al. consider a large-scale wireless network that uses sensors along its edge to estimate the characteristics of interference from neighboring networks or devices. The authors propose and justify a low-complexity threshold design technique in which the sensors use nonidentical thresholds to generate their bits. This technique produces a dithering effect that provides better performance than previous techniques that use different nonidentical thresholds or the case in which all the sensor motes use an identical nonoptimal threshold.

The energy efficiency optimization of the binary power control scheme for MIMO-OFDM wireless communication systems is formulated by Y. Wang et al. in the paper "*Modeling and performance analysis of energy efficiency binary power control in MIMO-OFDM wireless communication systems,*" wherein a global optimization solution of power allocation is also derived. Furthermore, a new energy efficiency binary power control algorithm is designed to improve the energy

efficiency of MIMO-OFDM wireless communication systems.

In the survey paper "*A survey of adaptive and real-time protocols based on IEEE 802.15.4*," F. Xia et al. discuss the negative aspects of IEEE 802.15.4 MAC in contention access period, contention-free period, and the overall cross-period, respectively, in terms of adaptive and real-time guarantees. They then give an overview on some interesting mechanisms used in existing adaptive and real-time protocols in compliance with IEEE 802.15.4.

Underwater acoustic sensor networks (UASNs) are playing a key role in ocean applications. The paper "*A reliable and efficient MAC protocol for underwater acoustic sensor networks*" by J. Xiong et al. proposes a MAC protocol called RAS, a priority scheduling approach for multihop UASNs, which is more efficient in throughput and delay performance. Furthermore, a reliable RAS called RRAS that obtains a tradeoff between the reliability and the efficiency is also suggested.

Design of MAC protocols supporting multimedia traffic QoS (Quality of Service) is challenging. In the paper "*A scalable MAC protocol supporting simple multimedia traffic QoS in WSNs*," J. On et al. propose a scalable MAC protocol that guarantees multimedia traffic, still images, and scalar sensor data QoS in multihop WSNs. The proposed MAC protocol outperforms the IEEE 802.11e EDCF and the IEEE 802.15.4 MAC protocol in terms of the end-to-end delay and stable transmission of multimedia streaming data.

The paper "*Quadratic programming for TDMA scheduling in wireless sensor networks*" is concerned with developing a novel protocol to achieve energy-efficient and reliable multihop data transfer in WSNs satisfying given latency requirements. The authors present a novel multihop aperiodic scheduling (MAS) algorithm that guarantees energy-efficient data collection by WSNs under delay constraints.

In the paper "*Link prediction and route selection based on channel state detection in UASNs,*" J. Chen et al. propose a model to predict link interruption and route interruption in UASNs by the historical link information and channel state obtained by periodic detection. A method of decomposing and recomposing routes hop by hop in order to optimize route reselection is also presented. Moreover, the authors present a back-up route maintenance scheme to keep back-up routes with fresh information.

The paper "*SEF: a secure, efficient, and flexible range query scheme in two-tiered sensor networks*" considers large-scale WSNs following the two-tiered architecture. The authors present a novel scheme called SEF for secure range queries. To preserve privacy, SEF employs the order preserving symmetric encryption which not only supports efficient range queries but also maintains a strong security standard. To preserve authenticity and integrity of query results, the authors propose a novel data structure called Authenticity & Integrity tree.

In the context of a body area network (BAN), the confidentiality and integrity of the sensitive health information is particularly important. In the paper "*A biometric key establishment protocol for body area networks*," L. Yao et al. present an electrocardiogram- (ECG-) signal-based key establishment protocol to secure the communication between every sensor and the control unit before the physiological data is transferred to external networks for remote analysis or diagnosis.

In mission-critical CPS, dependability is an important requirement at all layers of the system architecture. In the paper "*Anonymous aggregator election and data aggregation in wireless sensor networks*," T. Holczer and L. Buttyán propose protocols that increase the dependability of WSNs. More specifically, the authors propose two private aggregator node election protocols, a private data aggregation protocol, and a corresponding private query protocol for sensor networks that allow for secure in-network data aggregation by making it difficult for an adversary to identify and then physically disable the designated aggregator nodes.

In the paper "*A mobile-beacon-assisted sensor network localization based on RSS and connectivity observations*," G. Teng et al. propose a distributed mobile-beacon-assisted localization scheme based on received signal strength (RSS) and connectivity observations (named MRC) with a specific trajectory in static WSNs. To meet the requirements of real applications in practice, the authors develop two improved approaches based on MRC to consider irregular radio scenario in noisy environments.

In the paper "*Total least squares method for robust source localization in sensor networks using TDOA measurements*," the TDOA-based source localization problem in sensor networks is considered with sensor node location uncertainty. A total least squares (TLSs) algorithm is developed by a linear closed-form solution for this problem, in which the uncertainty of the sensor location is formulated as a perturbation.

Smart phones have become useful tools to implement high-confidence CPS. Context sharing systems in mobile environment attract attention with the popularization of social media. In the paper "*Bayesian network-based high-level context recognition for mobile context sharing in cyber-physical system*", H.-S. Park et al. propose a mobile context sharing system that can recognize high-level contexts automatically by using Bayesian networks based on mobile logs. They have developed a ContextViewer application which consists of a phonebook and a map browser to show the feasibility of the system.

The paper "*Distributed algorithm for traffic data collection and data quality analysis based on wireless sensor networks*" by N. Ding et al. deals with urban traffic data collection. The authors propose a distributed algorithm to collect the traffic data based on sensor networks and improve the reliability of data by quality analysis. The performance of this algorithm is analyzed with a number of simulations based on data set obtained in urban roadway.

Acknowledgments

It has been a great pleasure for us to organize this special issue. We would like to thank Professor Sundaraja Sitharama Iyengar, Editor-in-Chief of International Journal

of Distributed Sensor Networks (IJDSN), for giving us the opportunity. We also thank all authors for their submissions and all reviewers for their diligent work in evaluating these submissions. We sincerely hope that you enjoy reading these distinguished papers.

Feng Xia
Tridib Mukherjee
Yan Zhang
Ye-Qiong Song

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 329524, 12 pages
doi:10.1155/2011/329524

Research Article

A Game Theoretic Approach for Interuser Interference Reduction in Body Sensor Networks

Guowei Wu, Jiankang Ren, Feng Xia, Lin Yao, Zichuan Xu, and Pengfei Shang

School of Software, Dalian University of Technology, Dalian 116620, China

Correspondence should be addressed to Feng Xia, f.xia@ieee.org

Received 27 March 2011; Accepted 30 June 2011

As a kind of small-scale cyber-physical systems (CPSs), body sensor networks (BSNs) can provide the pervasive, long-term, and real-time health monitoring. A high degree of quality-of-service (QoS) for BSN is extremely required to meet some critical services. Interuser interference between different BSNs can cause unreliable critical data transmission and high bite error rate. In this paper, a game theoretic decentralized interuser interference reduction scheme for BSN is proposed. The selection of the channel and transmission power is modeled as a noncooperative game between different BSNs congregating in the same area. Each BSN measures the interference from other BSNs and then can adaptively select the suitable channel and transmission power by utilizing no-regret learning algorithm. The correctness and effectiveness of our proposed scheme are theoretically proved, and the extensive experimental results demonstrate that the effect of inter-user interference can be reduced effectively with low power consumption.

1. Introduction

The recent significant advances in wireless biomedical sensors, ubiquitous computing, and wireless communication offer great potential for the body sensor networks (BSNs) that are small-scale cyber-physical systems (CPSs). BSNs are becoming increasingly common in enabling many of the pervasive computing technologies that are becoming available today such as smart-homes and pervasive health monitoring systems [1–5]. By using various wireless wearable or implanted sensors which are capable of sensing, processing, and communicating one or more vital signs, BSNs can continuously monitor physiological vital signs (such as EEG, ECG, EMG, blood pressure, and oxygen saturation), physical activities (such as posture and gait) and environmental parameters (such as ambient temperature, humidity, and presence of allergens) and transmit them in real time to a centralized location for remote diagnosis. BSNs bring a significant impact to the scope of medical services in areas such as rehabilitation, geriatric care, sports medicine, fall detection, and gait analysis.

The challenges faced by BSNs are in many ways similar to general wireless sensor networks (WSNs), but there are intrinsic differences between the two which require special attention. Reliability is one of the crucial elements of BSN, as sensitive health information is being transmitted through the wireless [6]. However, compared with the conventional WSN, the architecture for BSN is highly dynamic, placing more rigorous constraints on power supply, communication bandwidth, storage, and computational resources. Different from the general WSN, as a human-centered sensor network, BSNs are usually deployed in open environments and the mobility of BSN users implies that individuals carrying BSN may meet other BSN users, and thus radio transmissions on different BSN users will interfere with each other when BSN users gather in the same place and their sensors transmit data simultaneously. Furthermore, such interuser interference will be increased in some architecture which predisposes some of the sensor nodes to transmit with high power. In addition, some extraneous interference in the same band (e.g., WiFi and bluetooth) can also increase the interuser interference [7]. Interuser interference causes unreliable critical data transmission and high bite error rate; hence, the critical data cannot be sent to control nodes in time and the quality of health monitoring is affected directly. Consequently, the doctor may give the wrong diagnosis and the patient may be delayed to cure and even die. Moreover, the packet loss caused by the interuser interference results in data retransmission. Due to the characteristics of energy sensitiveness and resource limitedness, numerous data retransmissions consume much energy and thus may

impel the node to fail. Thus, the interuser interference in BSN should be overcome before successful deployment. In the literature, several mechanisms (outlined in Section 2) have been proposed to mitigate the problems of interference in wireless sensor networks. However, there is almost no underlying interuser interference reduction infrastructure particularly for BSNs. To this end, our work aims to design an interuser interference reduction scheme for BSN.

In this paper, we propose a game theoretic decentralized interuser interference reduction scheme, namely DISG, to enhance the reliability in BSN. In our previous work [8], we briefly discussed the advantages of the proposed scheme. In this paper, we give more comprehensive analysis and extensive simulations. The key contribution of our work is summarized as follows. Firstly, game theory is used as a mathematical basis for the analysis of the interactive decision making process between the rational BSN users. The selection of the channel and transmission power is modeled as a noncooperative game between different BSN users congregating in the same area to reduce the effect of interuser interference. Thus, DISG is distributed, noncooperative, and self-organizing. In the scheme, each BSN just needs to measure the interference from other BSNs and then adaptively select the suitable channel and transmission power by utilizing no-regret learning algorithm without intercommunication. Secondly, different from the relative researches, the deployment of the fixed WSN infrastructure is not required in DISG. Therefore, the system with DISG is more flexible and suitable for outdoor environments. Finally, DISG does not only focus on the reliability but trades off the signal to interference-plus-noise ratio (SINR) and the energy cost as well. Moreover, the weighting factors can be configured dynamically based on the remaining battery capacity to adapt to different situations. For that reason, this scheme can reduce the effect of interuser interference with low power consumption. In addition, the feasibility proof and performance analysis of the DISG are presented. Note that our scheme could be applied in a variety of scenarios, for example, hospitals, geriatric care, sports, rehabilitation, and battlefield. However, for simplicity we use the terminology patient, physician, and medical service in this paper.

The remainder of the paper is organized as follows. In Section 2, we summarize the related work and describe the key limitations of them. Subsequently, the system architecture and the definitions used in the DISG scheme are given in Section 3. Section 4 presents the detailed description of the proposed game theoretic framework in the DISG scheme. In Section 5, the simulation and performance evaluation are presented. Finally, we conclude the paper in Section 6.

2. Related Works

Extensive research has been focusing on the interference analysis and reduction for the conventional wireless networks. Reference [9] relaxes the assumption that the interference range is equal to the reception range and develops closed-form expressions for the frequency of interference occurrence. Reference [10] studies the convergence of the average consensus algorithm in wireless networks in the presence of interference and characterizes the convergence properties of an optimal Time Division Multiple Access protocol that maximises the speed of convergence on these networks. Reference [11] addresses the problem of interference modeling for wireless networks and analyzes standard interference functions and general interference functions. Reference [12] studies the minimum-latency broadcast scheduling problem in the probabilistic model and establishes an explicit relationship between the tolerated transmission-failure probability and the latency of the corresponding broadcast schedule. Reference [13] investigates the impact of time diversity and space-time diversity schemes on the traffic capacity of large-scale wireless networks with low-cost radios and analyzes the tradeoff between relaying gain and increased interference due to additional traffic. Reference [14] establishes a fundamental lower bound on the performance of a wireless system with single-hop traffic and general interference constraints. Reference [15] proposes a novel consensus algorithm based on the simple self-synchronization mechanism in Reference [16], which is effective in suppressing both noise and interference for arbitrary network topology. In [17], the authors propose a cluster-based MAC protocol combining TDMA, FDMA with CSMA/CA, which adopts TDMA mechanism in the clusters and FDMA mechanism among different clusters to reduce the interference. Reference [18] presents a self-reorganizing slot allocation (SRSA) mechanism for TDMA-based medium access control (MAC) in multicluster sensor networks to reduce inter-cluster TDMA interference without having to use spectrum expensive and complex wideband mechanisms. Reference [19] proposes an algorithm named "Minimizing Interference in Sensor Network (MI-S)" to minimize the maximum interference for set of nodes in polynomial time maintaining the connectivity of the graph. Reference [20] presents a new greedy heuristic channel assignment algorithm (termed CLICA) for finding connected, low interference topologies. References [21, 22] treat the problem of assigning power level to a set of nodes in the plane as the problem of yielding a connected geometric graph and study the performance of a number of heuristics through simulation to minimize interference in wireless sensor networks. However, there is a significant gap between the conventional wireless sensor network systems due to the characteristics of BSN, especially, the mobility of BSN, and thus none of these interference reduction mechanisms in WSN can be directly applied in BSN.

The reliability research of BSN mainly focuses on the reliability of intra-BSN communications. In [23, 24], a novel QoS cross-layer scheduling mechanism based on fuzzy-logic rules is proposed. An energy-saving distributed queuing MAC protocol is adopted. It can guarantee all packets are served with a specific bit-error-rate (BER) and within the particular latency limit while keeping low power consumption. In [25, 26], an active replication algorithm for the reliability of communication in multihop BSN is proposed, which can guarantee k-connectivity between nodes. In [27], the challenges brought by BSN applications are presented and a statistical bandwidth strategy, namely BodyQoS, is proposed to guarantee reliable data

communication. Reference [28] presents an adaptive and flexible fault-tolerant communication scheme for BSNs to fulfill the reliability requirements of critical sensors by a channel bandwidth reservation strategy based on the fault-tolerant priority and queue. In [29], an interference-aware topology control algorithm is proposed for wireless personal area network (WPAN) to minimize the interference effects by adjusting the transmission power dynamically based on the interference effect and maintaining the local topology of each sensor node adaptively. A new quality of service (QoS) routing protocol that relies on traffic diversity of these applications and ensures a differentiation routing using QoS metrics is proposed in [30]. Reference [31] presents a data-centric multiobjective QoS-aware routing protocol that facilitates the system to achieve customized QoS services for each traffic category differentiated according to the generated data types. Nonetheless, all of them neglect the problem of interuser interference. Reference [7] highlights the existence of the interuser interference effect in BSN from the perspective of network architectures and describes the interuser interference as the interference due to the communications of wireless BSNs operating in proximity of one another, but the issue of interuser interference is not investigated. In [32], the authors conduct a preliminary investigation of the impact and significance of interuser interference and investigate its behavior with respect to parameters. It gives three techniques that can be used to mitigate interference and implements an instance with a fixed WSN infrastructure to identify when BSNs are interfering with each other and make the BSNs communicate on different frequency channels to reduce interference between them. However, that paper does not conduct a comprehensive study of interuser interference and mainly intends to provide an indication of the impact of the interuser interference effect. In addition, the fixed WSN infrastructure is only suitable for indoor environments and thus limits the application of BSN. In [8] the decentralized interuser interference suppression approach with noncooperative game is briefly discussed and evaluated. As an extension of [8], in this paper we give the architecture sketch of the interference reduction system implemented on the control node of BSN and propose a weighing factor configuration algorithm to configure the weighting factors dynamically to adapt to different situations. In addition, the concepts of DISG are investigated more thoroughly, and the simulations are more elaborated.

3. System Architecture and Definitions

In this section, we present the system architecture and the definitions used in the DISG scheme. Figure 1 shows a generalized overview of the system architecture that encompasses a set of intelligent physiological sensors, control nodes (Internet-enabled PDA or cell phone), and base stations and a network of remote health care servers and related services (such as caregiver, physician). In general, the biosensor is a wireless wearable or implanted vital sign sensor (such as sweat, EKG, and temperature) that consists of a processor, memory, transceiver, sensors, and a power unit. Each biosensor node is responsible for sensing one or more physiological signals, processing these signals (coding, filtering, aggregation, feature extraction, feature recognition, etc.), storing the processed data, and forwarding the data to the control node. Due to the size and energy consumption restrictions, the biosensor has limited memory and power and cannot afford heavy computation and communication. Compared to the biosensor, the control node that is a PDA or a 3G cell phone has relatively high energy and computing capability. In other words, BSN is a typical asymmetric structure. The control node provides the human-computer interface and communicates with the remote medical server(s) through the base station to result in patient-specific recommendations [33]. It is worth mentioning that the above system architecture is not only suitable for medical healthcare applications but also for sports and battlefield.

As illustrated in Figure 1, the transmissions of sensors in different users will interfere with each other, when they operate in the same vicinity and communicate concurrently. In order to reduce the effect of interuser interference in the network, the DISG scheme is proposed to guarantee reliability performance by modeling the problem as a noncooperative game to decentralize the interuser interference reduction. We first give some definitions in the DISG scheme, as listed in Table 1.

Definition 1. Interuser interference: interuser interference is the interference in desired communication when several BSNs operate in the same vicinity.

Definition 2. Noncooperative interuser interference reduction game: the noncooperative game is made up of three basic components, a set of players, a set of actions, and a set of costs, and represented by the tuple $\langle \mathrm{B}, \Lambda, \mathrm{U} \rangle$.

Definition 3. Players in the system: in this game, each BSN user is a player as the decision maker in the modelled scenario, and $\mathrm{B} = \{b_i, i = 1, \ldots, M\}$ denotes the set of players in the system and M is the number of users.

Definition 4. Available nonoverlapping frequency channels: the set of all available nonoverlapping frequency channels for BSN users is represented by $\mathrm{C} = \{c_j, j = 1, \ldots, N\}$, where N is the number of the available nonoverlapping channels.

Definition 5. The action space of the system: $\Lambda = \Lambda_1 \times \cdots \times \Lambda_M$ represents the action space of the system. The strategy space of user b_i is defined as $\Lambda_i = \{\sigma_i := \{\rho_i^j p_i^j\} \mid j \in \{1, \ldots, N\}\}$, where we define the mixed strategy as $\rho_i^j \in \{0, 1\}$, $\sum_{j=1}^{N} \rho_i^j = 1$, and $\rho_i^j = 1$ indicates that b_i selects the channel c_j, which means, in this distribution, that only the pure strategy ρ_i^j is assumed to be selected with a probability of 1. p_i^j represents the transmission power of b_i on the channel c_j and $p_i^j \in [p_{\min}, p_{\max}]$, where $p_{\min}$ and $p_{\max}$ are the minimum transmitted power and the maximum transmitted power, respectively, due to the characteristic of the specific physiological sensors.

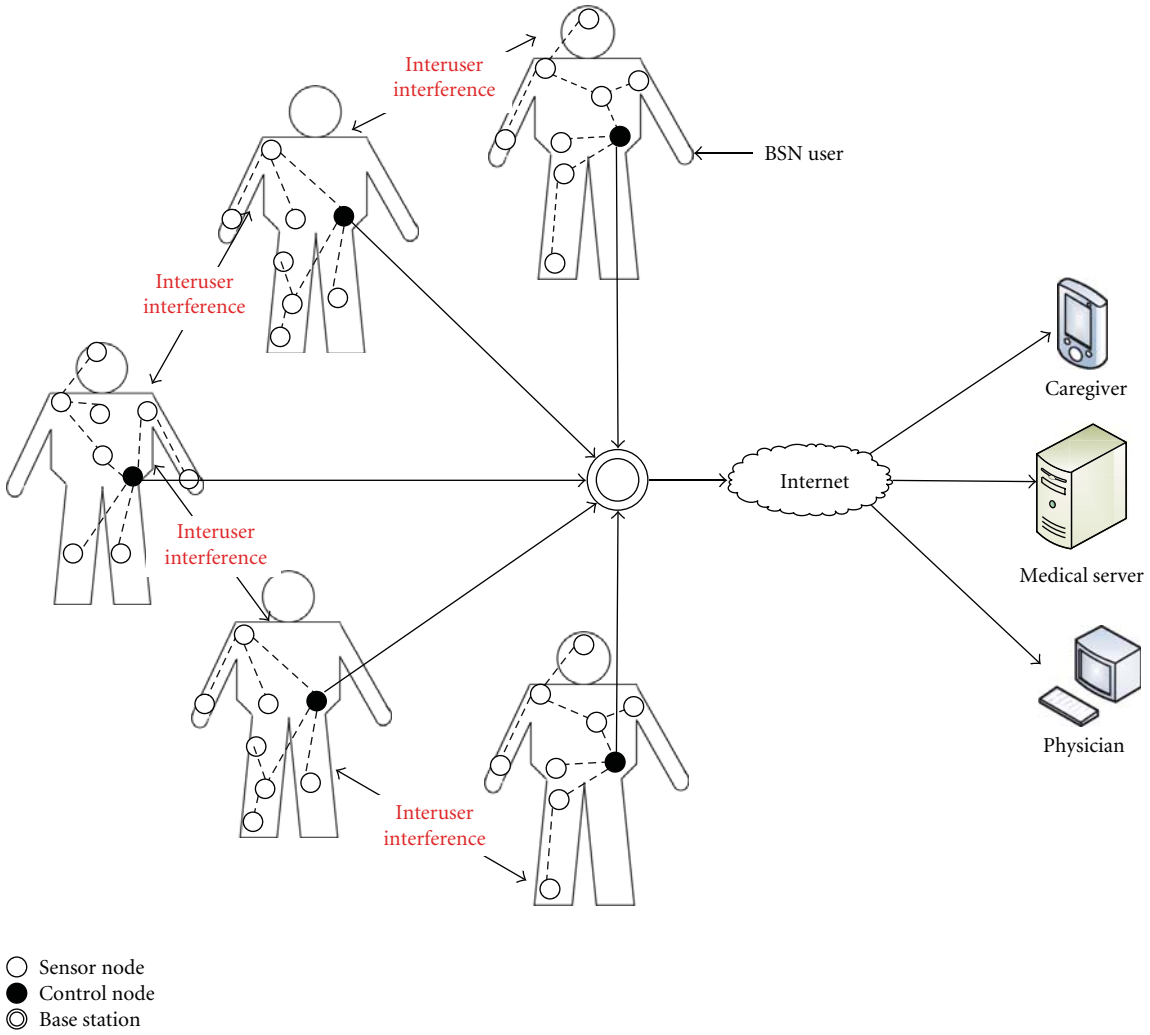


Figure 1: System architecture.

Definition 6. The cost set of the system: the cost set of the system is denoted by $U = \{u_1, \ldots, u_M\}$. The cost of user b_i is denoted as $u_i(\sigma_i, \sigma_{-i})$, which is a function of all users' actions. Each user selects its own action σ_i to minimize the cost function. σ_{-i} is the action taken by users except b_i.

4. Interuser Interference Reduction Game Theoretic Framework

Figure 2 illustrates the architecture sketch of the proposed framework that is implemented on the control node of BSN. This framework consists of four components: interference detector, strategy analyzer, strategy selector, and strategy actuator. The interference detector monitors interference plus noise and produces periodic reports. The strategy analyzer examines the report derived from the interference detector and the available channel as well as transmission power level. Then the strategy selector uses the noncooperative game model between different BSNs (users) defined in the scheme to choose the suitable strategy. In the game model, each BSN user is selfish and tends to close to the ideal SINR to ensure specific QoS requirement without regard to the QoS of other BSN users. In the strategy selector component, the status of the current strategy is read in and checked against the QoS requirement. If the requirement is satisfied, no change is necessary and the strategy selector terminates. Otherwise, the rules defined in the game model are followed to determine the appropriate channel and transmission power level by utilizing no-regret learning algorithm, and the strategy decision is made accordingly. It is worth mentioning that any environment change such as user movement or noise variation might trigger a strategy adjustment. Finally, the strategy actuator will send the new strategy decision to sensor nodes deployed in the same BSN users to reconfigure RF to switch to the newly selected channel and uses the appropriate transmission power level. In the scheme, the interuser interference reduction game model is the key to making strategy decisions and will be discussed in detail in the following section.

Table 1: Variables and notations.

Variable	Description
$B = \{b_i, i = 1, \ldots, M\}$	Set of all BSN users
$C = \{c_j, j = 1, \ldots, N\}$	Set of nonoverlapping frequency channels
p_i	Transmission power of user b_i
G_i	The link gain of b_i
S_i	Attenuation coefficient of b_i
d_i	Maximum distance between sensor and control node in b_i
δ	Constant parameter modeling the shadowing effect
p_i^j	The transmission power of b_i on the channel c_j
v_k^i	The distance between user b_k and user b_i
η_i	The background noise power
γ_i^j	The SINR of b_i on the channel c_j
γ_i^0	The ideal SINR threshold of b_i
Λ	The action space of the system
U	The cost set of the system
σ_i	The action taken by b_i
σ_{-i}	Action taken by users except b_i
$u_i(\sigma_i, \sigma_{-i})$	The cost of b_i
$I_i = \{I_i^j(\sigma_{-i}), \forall c_j\}$	The local information about interference observed by b_i
π_i	The strategy for the selection of the channel and the transmitted power taken by b_i
ca_i	The maximum battery capacity of sensor deployed in b_i
$\rho^* p^*$	The Nash Equilibrium of the game
ω_i	The probability distribution of b_i over the set of strategies
$c_{j_i^*}$	The best channel selected by b_i
$\varphi \in \{1, \ldots, t\}$	The iteration period in the game
${}^{\varphi}r_i(\sigma_i(p_i^j)^*)$	The regret of b_i for action $\sigma_i(p_i^j)^*$ at iteration φ
${}^{t}D_i(\sigma_i(p_i^j)^*)$	The historical cumulative regret of b_i for action $\sigma_i(p_i^j)^*$
${}^{t+1}\omega_i(\sigma_i(p_i^j)^*)$	The probability of b_i for action $\sigma_i(p_i^j)^*$ at iteration t+1
${}^{t+1}\rho_i(\sigma_i((p_i^j)^*))$	The best strategy of b_i at iteration t+1

4.1. Cost Function. In the purpose of reducing the effect of interuser interference in the network, the selection of the channel and transmission power is modeled as a noncooperative game between different BSNs (users). In this game, each user is a rational player. They observe local conditions and neighbors' actions and independently adapt their strategy using no-regret learning algorithm to minimize the interuser interference. Obviously, this framework of the noncooperative game is propitious to solving the problem of the interuser interference in BSN. In the noncooperative game, the cost function normally represents the hate of a selfish user, who wants to minimize its own loss. Similarly,

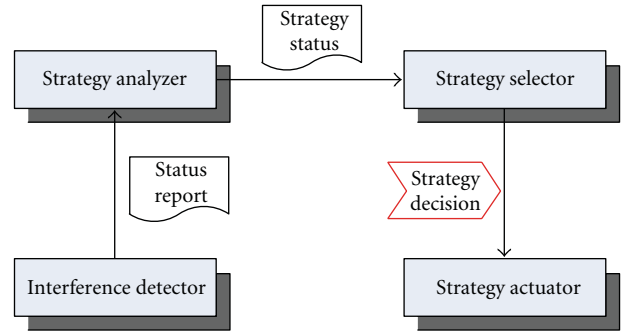


Figure 2: Interference reduction system on control node.

a player in our game may somehow modify its channel and transmission power in such a way that can not only decrease its own loss but also increase interferences to others. Due to the energy limitedness of the control node, the desired cost function should take account of the SINR and power utilization. Thus, similar to [34, 35], the cost function of each user that considers the tradeoff between attaining high SINR and maintaining lower power utilization is expressed as

$$u_i(\sigma_i, \sigma_{-i}) = \tau_i\left(\gamma_i^0 - \gamma_i^j\right)^2 + \xi_i p_i^j, \quad \rho_i^j = 1, \tag{1}$$

where τ_i and ξ_i are two nonnegative weighting factors. Different levels of emphasis on SINR and power utilization are obtained by adjusting the weighting factors. Choosing $\tau_i/\xi_i > 1$ places more emphasis on the SINR, whereas $\tau_i/\xi_i < 1$ puts more emphasis on the power usage. In the subsequent section, a weighing factor configuration algorithm will be presented to configure them dynamically to adapt to different situations. γ_i^0 is the ideal signal-to-interference-plus-noise ratio (SINR) threshold to which selfish user b_i wants to close to ensure specific QoS requirement, but whether other users meet their QoS requirements is irrelevant. γ_i^j is the SINR of b_i on the channel c_j, and, based on [36], it is calculated as follows:

$$\gamma_i^j = \frac{G_i p_i^j}{\sum_{k=1, k\neq i}^{M}\left(S_k p_k^j / \left(v_k^i\right)^{\delta}\right) + \eta_i}, \tag{2}$$

where p_i^j is the transmission power of b_i on the channel c_j, v_k^i is the distance between different users b_k and b_i, and η_i is the background noise power. $G_i p_i^j$ expresses the received power of user b_i, and $G_i > 0$ is the link gain of b_i and can be calculated as follows:

$$G_i = \frac{S_i}{d_i^{\delta}}, \tag{3}$$

where S_i is the attenuation coefficient, d_i is the maximum distance between the sensor node and the control node in the user b_i, and δ is a constant parameter modeling the shadowing effect.

Formally, based on (1) and (2), the cost function is defined as follows:

$$u_i(\sigma_i, \sigma_{-i}) = \tau_i \left(\gamma_i^0 - \frac{G_i p_i^j}{\sum_{k=1,k\neq i,\rho_k^j=1}^{M} \left(S_k p_k^j / \left(v_k^i \right)^\delta \right) + \eta_i} \right)^2 + \xi_i p_i^{j_i}, \quad \rho_i^j = 1. \tag{4}$$

In the game, the goal of user b_i is to attain high SINR and minimize power utilization over the channel that it chooses. It is worthwhile to note that there is not message exchange in the game. The cost distribution of other users σ_{-i} is not available for user b_i; however, b_i can get the aggregate response $I_i^j(\sigma_{-i})$ that is, the interference from other users plus noise. To perform (5), b_i needs to observe the local information $I_i = \{I_i^j(\sigma_{-i}), \text{ for all } c_j\}$. Based on the information, the following best response is adopted by every user in the system:

$$\pi_i(I_i) = \arg\min_{\sigma_i \in \Lambda_i} u_i(\sigma_i, I_i) \quad \forall b_i \in \mathrm{B}, \tag{5}$$

where π_i represents the strategy for the selection of channel and transmission power. The solution to the problem in (5) will lead to a Nash Equilibrium (NE).

4.2. Nash Equilibrium in the Game. In order to analyze the outcome of the game, we will give how the game will be played by rational players using the concept of Nash Equilibrium (NE), which states that at equilibrium every player selects a cost-minimizing strategy given the strategies of other players.

The Nash Equilibrium $\rho^* p^* = [\rho_1^{j_1^*} (p_1^{j_1^*})^*, \rho_2^{j_2^*} (p_2^{j_2^*})^*, \ldots, \rho_N^{j_N^*} (p_N^{j_N^*})^*]$ in the game is mathematically defined as

$$u_i\left(\rho_i^{j_i^*} \left(p_i^{j_i^*}\right)^*, \rho_{-i}^j p_{-i}^j\right) \leq u_i\left(\rho_i' p_i', \rho_{-i}^j p_{-i}^j\right), \quad \forall b_i \in \mathrm{B}, \forall \rho_i' p_i' \in \Lambda_i, \tag{6}$$

that is, given that other players' strategies remain unchanged, no player would be tempted to deviate from its current strategy to reduce its cost alone.

Theorem 1. *For a specific channel c_j, with all other users' transmission powers fixed on this channel, user b_i can get the best transmission power on the channel c_j, which is $(p_i^{j_i})^* = \arg\min_{p_i^{j_i} \in p_i} u_i(p_i^{j_i}, p_{-i}^{j_i})$ in the condition of $\tau_i/\xi_i \geq 2p_{\max}/(\gamma_i^0)^2$ and $\tau_i/\xi_i > I_i^{j_i}(p_{-i}^{j_i})/\gamma_i^0 G_i$.*

Proof. Let $\sum_{k=1,k\neq i,j_k\neq j_i}^{M} S_k p_k^j/(v_k^i)^\delta + \eta_i = I_i^{j_i}(p_{-i}^{j_i})$ so (4) can be written as

$$u_i\left(p_i^{j_i}, p_{-i}^{j_i}\right) = \tau_i \left(\gamma_i^0 - \frac{G_i p_i^{j_i}}{I_i^{j_i}\left(p_{-i}^{j_i}\right)} \right)^2 + \xi_i p_i^{j_i}, \quad \rho_i^{j_i} = 1. \tag{7}$$

Taking the derivative of (7) with respect to $p_i^{j_i}$ and equating it to zero for $p_i^{j_i}$ gives the following condition:

$$2\tau_i \left(\gamma_i^0 - \frac{G_i p_i^{j_i}}{I_i^{j_i}\left(p_{-i}^{j_i}\right)} \right) \left(-\frac{G_i}{I_i^{j_i}\left(p_{-i}^{j_i}\right)} \right) + \xi_i = 0, \tag{8}$$

that is,

$$\frac{2\tau_i G_i^2 p_i^{j_i}}{\left(I_i^{j_i}\left(p_{-i}^{j_i}\right)\right)^2} - \frac{2\tau_i \gamma_i^0 G_i}{I_i^{j_i}\left(p_{-i}^{j_i}\right)} + \xi_i = 0. \tag{9}$$

Thus, we can get

$$p_i^{j_i} = \frac{\gamma_i^0 I_i^{j_i}\left(p_{-i}^{j_i}\right)}{G_i} - \frac{\xi_i \left(I_i^{j_i}\left(p_{-i}^{j_i}\right)\right)^2}{2\tau_i G_i^2}. \tag{10}$$

Due to $p_i^{j_i} \geq 0$, we can get

$$\frac{\gamma_i^0 I_i^{j_i}\left(p_{-i}^{j_i}\right)}{G_i} - \frac{\xi_i \left(I_i^{j_i}\left(p_{-i}^{j_i}\right)\right)^2}{2\tau_i G_i^2} \geq 0, \tag{11}$$

that is,

$$\frac{\tau_i}{\xi_i} \geq \frac{I_i^{j_i}\left(p_{-i}^{j_i}\right)}{2\gamma_i^0 G_i}. \tag{12}$$

For given values of τ_i, ξ_i, γ_i^0, G_i, and $p_i^{j_i}$, the quadratic (10) in $I_i^{j_i}(p_{-i}^{j_i})$ has a real solution if and only if

$$\left(\frac{\gamma_i^0}{G_i}\right)^2 - 4\left(\frac{\xi_i}{2\tau_i G_i^2}\right) p_i^{j_i} \geq 0, \tag{13}$$

that is,

$$p_i^{j_i} \leq \frac{\tau_i \left(\gamma_i^0\right)^2}{2\xi_i}. \tag{14}$$

Because the transmission power is $p_i^j \in [p_{\min}, p_{\max}]$, based on inequality (14), τ_i/ξ_i must satisfy

$$\frac{\tau_i}{\xi_i} \geq \frac{2p_{\max}}{\left(\gamma_i^0\right)^2}. \tag{15}$$

According to [36], in order to make the game converge to a unique fixed point, a fixed point from (10) should satisfy the following properties: (1) positivity; (2) monotonicity; (3) scalability. Therefore, based on [34], we can get

$$\frac{\tau_i}{\xi_i} > \frac{I_i^{j_i}\left(p_{-i}^{j_i}\right)}{\gamma_i^0 G_i}. \tag{16}$$

Based on (12), (15), and (16), we can get that the noncooperative game has a unique NE for the transmission power $(p_i^{j_i})^*$ on the channel c_j in the condition of (15) and (16). □

Require: γ_i^0, G_i, $p_{\max}$, ca_i, α_i, β_i
Ensure: τ_i, ξ_i
(1) $\xi_i = 1$
(2) Estimate $I_i^{j_i}(p_{-i}^{j_i})$ associated to the current channel and transmission power.
(3) $temp = \mathrm{Max}(2 \times p_{\max}/(\gamma_i^0)^2,\ I_i^{j_i}(p_{-i}^{j_i})/(\gamma_i^0 G_i))$
(4) Obtain the set of the remaining battery capacity of each sensor $CA_{remaining}$.
(5) $\mathrm{ca}_{\min} = \mathrm{getMin}(CA_{remaining})$
(6) **for** $(i = 0, i \le \mu, i++)$
(7) **if** $(\mathrm{ca}_{\min} \ge \mathrm{ca}_i \times ((\mu - i)/\mu))$
(8) $\tau_i = temp + \alpha_i + (\mu - i) \times \beta_i$
(9) **return**
(10) **end if**
(11) **end for**

Algorithm 1: Weighing factor configuration algorithm.

In order to configure the weighting factors τ_i and ξ_i dynamically to adapt to different situations, we propose a weighing factor configuration algorithm based on Theorem 1 as shown in Algorithm 1. In the scheme, the control node collects the minimum remaining battery capacity $\mathrm{ca}_{\min}$ of sensor nodes deployed in the same BSN and observes the local information $I_i^{j_i}(p_{-i}^{j_i})$. Then, the relation between τ_i and ξ_i can be determined based on $I_i^{j_i}(p_{-i}^{j_i})$ according to the relation between the minimum remaining battery capacity $\mathrm{ca}_{\min}$ and the maximum battery capacity ca_i. It is worth noting that the maximum battery capacity of each sensor deployed in the same BSN is assumed to be the same. In the algorithm, α_i and β_i are two negative constant parameters. We can find that the greater the minimum remaining battery capacity is, the greater the value of τ_i/ξ_i is.

A probability distribution ω is a correlated equilibrium (CE) over the set of user strategies if and only if the following inequality is satisfied:

$$\sum_{\rho_{-i} \in \Lambda_i} \omega(\rho_i, \rho_{-i}) [u_i(\rho_i', \rho_{-i}) - u_i(\rho_i, \rho_{-i})] \ge 0 \tag{17}$$

for all $b_i \in \mathrm{B}$, and $\rho_i \in \Lambda_i$ and $\rho_{-i} \in \Lambda_{-i}$, and a CE is an NE if and only if it is a product measure.

Theorem 2. *Each user b_i converges to an NE to attain the best channel $c_{j_i^*}$ (i.e., $\rho_i^{j_i^*} = 1$) with the no-regret learning approach.*

Proof. Based on (17), we can find that the set of correlated equilibrium is nonempty, convex, and closed in the finite game. Thus, according to [37], when players select their strategy randomly, there exists a no-regret strategy such that, if every player follows such a strategy, then the empirical frequencies of play converge to the set of correlated equilibrium. In addition, this correlated equilibrium is NE, because the correlated equilibrium corresponds to the special case in which $\omega(\rho_i, \rho_{-i})$ is a product of each player's probability for different actions, that is, the play of the different players is independent. Thus, based on the NE converged, each user b_i can attain the best channel $c_{j_i^*}$ through no-regret learning approach repeatedly. □

Theorem 3. *Nash Equilibrium exists in the noncooperative interuser interference reduction game.*

Proof. For each user $b_i \in \mathrm{B}$, it can attain the best channel $c_{j_i^*}$ (i.e., $\rho_i^{j_i^*} = 1$) with the no-regret learning approach based on Theorem 2 and select the best transmission power $p_i^{j_i^*}$ on the channel $c_{j_i^*}$ according to Theorem 1 to minimize cost. Therefore, there exists an NE $\rho^* p^* = [\rho_1^{j_1^*}(p_1^{j_1^*})^*, \rho_2^{j_2^*}(p_2^{j_2^*})^*, \ldots, \rho_N^{j_N^*}(p_N^{j_N^*})^*]$ in the game. □

4.3. No-Regret Learning Algorithm for the Game. In order to analyze the behaviour of the selfish users in the noncooperative interuser interference reduction game, we resort to the regret minimization learning algorithms [38]. No-regret learning algorithms are probabilistic learning strategies that specify that players explore the space of actions by playing all actions with some nonzero probability and exploit successful strategies by increasing their selection probability. In the game, each user b_i can determine the probability distribution of strategy σ_i based on the average regret at the iteration period $\varphi \in \{1, \ldots, t\}$, and thus it can select its action based on the history of actions of every user. The regret is defined as the difference between the costs caused by distinct strategies, and thus, for every two distinct strategy profiles $({}^{\varphi}\sigma_i, {}^{\varphi}\sigma_{-i})$ and $(\sigma_i((p_i^j)^*), {}^{\varphi}\sigma_{-i})$, user b_i uses the strategy profile $(\sigma_i((p_i^j)^*), {}^{\varphi}\sigma_{-i})$ to replace the strategy profile $({}^{\varphi}\sigma, {}^{\varphi}\sigma_{-i})$ and the user b_i feels a regret:

$${}^{\varphi}r_i\left(\sigma_i\left(p_i^j\right)^*\right) = u_i({}^{\varphi}\sigma) - u_i\left(\sigma_i\left(p_i^j\right)^*, {}^{\varphi}\sigma_{-i}\right). \tag{18}$$

The resulting historical cumulative regret of user b_i from $({}^{\varphi}\sigma_i, {}^{\varphi}\sigma_{-i})$ to $(\sigma_i((p_i^j)^*), {}^{\varphi}\sigma_{-i})$ up to iteration t is

$${}^{t}D_i\left(\sigma_i\left(p_i^j\right)^*\right) = \max\left\{\frac{1}{t}\sum_{\varphi=1}^{t}\left[u_i({}^{\varphi}\sigma) - u_i\left(\sigma_i\left(\left(p_i^j\right)^*\right), {}^{\varphi}\sigma_{-i}\right)\right], 0\right\}. \tag{19}$$

Require: $\mathrm{C} = \{c_j, j = 1, \ldots, N\}$, γ_i^0, τ_i, ξ_i, G_i, $p_{\min}$, $p_{\max}$
Ensure: The strategy with the suitable channel and suitable transmission power
(1) $\mathrm{C}_{selected} = \Phi$
(2) **while** ($\mathrm{C}_{selected} \neq \mathrm{C}$)
(3) c_{j_i} = random(C)
(4) $\mathrm{C}_{selected} = \mathrm{C}_{selected}$.add ($c_{j_i}$)
(5) Estimate the interference from other BSN users plus noise $I_i^j(\sigma_{-i})$
(6) Obtain the optimal transmission power $(p_i^{j_i})^*$ over channel c_{j_i} using (10)
(7) Get the cost $u_i(\sigma_i, \sigma_{-i})$ based on $(p_i^{j_i})^*$ and c_{j_i} using (7)
(8) Calculate the regret ${}^{\varphi}r_i(\sigma_i(p_i^j)^*)$ for two distinct strategies using (18)
(9) Get the historical cumulative regret ${}^{t}D_i(\sigma_i(p_i^j)^*)$ using (19)
(10) Update the probability of the channel selection ${}^{t+1}\omega_i(\sigma_i(p_i^j)^*)$ using (20)
(11) **end while**
(12) **while** ($\max(\omega_i) \leq \phi$)
(13) c_{j_i} = randomWithProbability(C, ω_i)
(14) Estimate the interference from other BSN users plus noise $I_i^j(\sigma_{-i})$
(15) Obtain the optimal transmitted power $(p_i^{j_i})^*$ over channel c_{j_i} using (10)
(16) Get the cost $u_i(\sigma_i, \sigma_{-i})$ based on $(p_i^{j_i})^*$ and c_{j_i} using (7)
(17) Calculate the regret ${}^{\varphi}r_i(\sigma_i(p_i^j)^*)$ for two distinct strategies using (18)
(18) Get the historical cumulative regret ${}^{t}D_i(\sigma_i(p_i^j)^*)$ using (19)
(19) Update the probability of the channel selection ${}^{t+1}\omega_i(\sigma_i(p_i^j)^*)$ using (20)
(20) **end while**
(21) Determine the strategy using (21)

Algorithm 2: No-regret learning algorithm for the game.

Based on (19), the probability ω_i assigned to the action $\sigma_i(p_i^j)^*$ of user b_i at iteration $t+1$ is calculated by

$$ {}^{t+1}\omega_i\left(\sigma_i\left(p_i^j\right)^*\right) = \frac{{}^{t}D_i\left(\sigma_i\left(\left(p_i^j\right)^*\right)\right)}{\sum_{\sigma_i(p_i^{j'})\in\Lambda_i} {}^{t}D_i\left(\sigma_i\left(\left(p_i^{j'}\right)^*\right)\right)}. \quad (20) $$

The above learning algorithm can converge to Nash equilibrium in the game, and thus each user b_i repeats calculations based on (19) and (20), and then it can determine its strategy at time step $t+1$ as

$$ {}^{t+1}\rho_i\left(\sigma_i\left(\left(p_i^j\right)^*\right)\right) = \begin{cases} 1 & \text{if } {}^{t+1}\ \omega_i\left(\sigma_i\left(\left(p_i^j\right)^*\right)\right) = \max\left\{{}^{t+1}\omega_i\right\}. \\ 0 & \text{otherwise} \end{cases} \quad (21) $$

By following the no-regret learning adaptation process, the user can learn how to choose the channel and transmission power to minimize the cost through repeated play of the game. Algorithm 2 presents the pseudocode of the iterative no-regret learning algorithm used in the interuser interference reduction game. At the initial step, each user b_i randomly selects its channel c_{j_i} with equal probability as the initial action. The optimal transmission power $p_i^{j_i}$ and cost $u_i(\sigma_i, \sigma_{-i})$ on the channel c_{j_i} will be calculated based on the interference plus noise measured. Based on $p_i^{j_i}$ and $u_i(\sigma_i, \sigma_{-i})$, the probability of b_i for the current strategy is updated accordingly. If there exists a channel that has not been selected randomly with equal probability, then user b_i repeats the step of random channel selection with equal probability. When the maximum $\omega_i(\sigma_i(p_i^j)^*)$ in the probability distribution ω_i has exceeded the threshold ϕ, user b_i will get the solution $\rho_i^{j_i^*}(p_i^{j_i^*})^*$, otherwise it will select a channel based on the probability distribution ω_i randomly and repeat its learning process in a round-robin manner to favour the strategy with minimum interference plus noise.

It is worth mentioning that the no-regret learning algorithm is a distributed iterative best-response scheme. In the noncooperative game, based on the information available locally, each player determines its own strategy (i.e., the channel and transmission power) to ensure that the solution converges. The convergence of the solution indicates that every user achieves its desirable operating point, where no user intends to change its strategy.

5. Simulation and Performance Analysis

5.1. Simulation Model and Method. The performance of the proposed interuser interference reduction scheme is evaluated by extending the Castalia simulator [39] to support the simulation of interuser interference in BSN. Castalia is an open source, discrete event-driven simulator designed specifically for WSN, BAN, and generally networks of low-power embedded devices based on OMNeT++ [40].

Table 2: Simulation parameters.

Parameters	Value
Number of BSNs	5
Area	10 m × 10 m
Number of sensor types	3 {ECG, SpO2, Temperature}
Number of available nonoverlapping frequency channels	5 {11, 12, 13, 14, 15}
Transmission power set (mW)	{29.04, 32.67, 36.3, 42.24, 46.2, 50.69, 55.18, 57.42}
Wireless channel model	Log shadowing wireless model
Path loss exponent	2.4
Collision model	Additive interference model
Physical and MAC layer	IEEE 802.15.4 standard
Application type	Event-driven

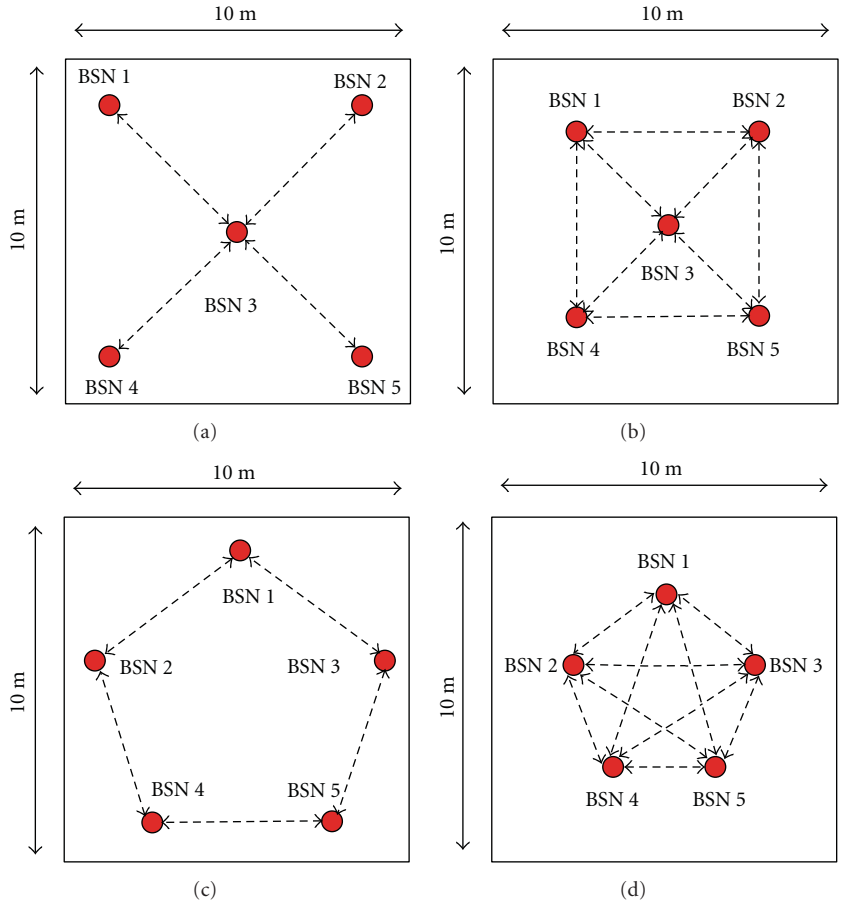


Figure 3: System topology.

For simulation purposes, five BSNs are distributed over a 10 m × 10 m square area, and each BSN simulates a typical on-body sensor network, in which sensors measure a person's physiological parameters. Each BSN has a control node and three biosensors (i.e., ECG sensor, SpO2 sensor, and temperature sensor). All sensor nodes adopt Castalia standard CC2420 IEEE802.15.4 radios. The number of available nonoverlapping frequency channels for each BSN is five in the experiment. The detailed configuration of simulation parameters is shown in Table 2. Note here that the configuration of parameters may vary depending upon the kind of sensors. In order to evaluate the performance of the scheme to copy with interuser interference for the different user densities and system typologies generated by the mobility of BSN users, we set up four representative system topologies to perform the simulation. Figure 3 presents the BSN user distributions of four representative system topologies during the simulation, where the dotted line

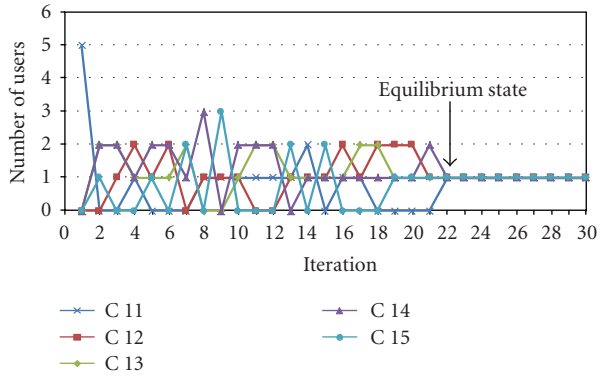


Figure 4: User distribution on different channels.

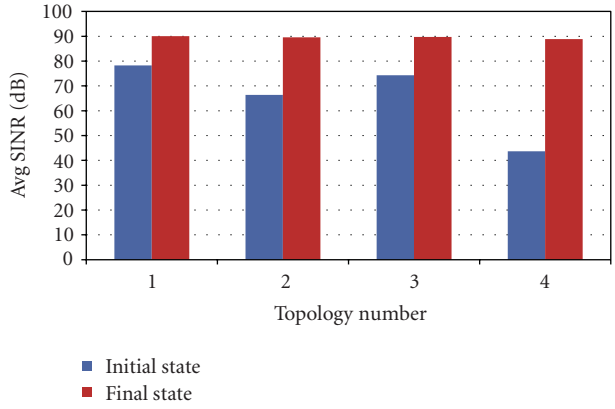


Figure 5: Average SINR initially and after convergence.

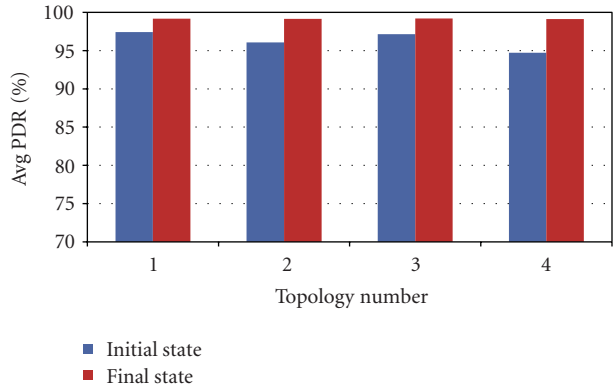


Figure 6: Average PDR initially and after convergence.

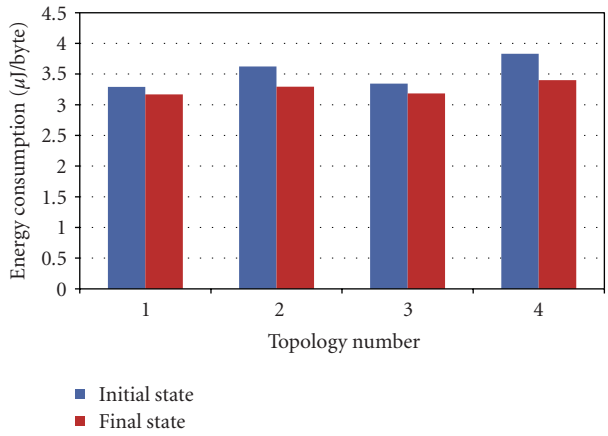


Figure 7: Energy consumption of per delivered byte initially and after convergence.

represents the interuser interference between two users. In the initial state, all BSN users use the same channel and select a random transmission power level as the initial strategy. In our experiment, we compare the performance of the system in the initial state and the final state after convergence to demonstrate the performance of the DISG scheme.

5.2. Simulation Results. We repeated the simulations several times and got similar results in each trial. In the following, we present a representative result from one trial and analyze it in detail.

The convergence property of the proposed no-regret learning algorithm for the interuser interference reduction game is evaluated first to demonstrate the correctness and effectiveness of the scheme. From Figure 4 that shows the user distribution on different channels over time during the simulation of the proposed no-regret learning algorithm for the topology 4, we can find that, at the beginning of the experiment, all BSN users select their channels randomly with equal probability, and, at last, the learning solution converges to a stable and desirable state (i.e., the equilibrium state) after a few iterations where all BSN users select different frequency channels to minimize interference.

As performance measures for the proposed interuser interference reduction scheme, we consider the average SINR of the system (Figure 5), the packet delivery rate (PDR) (Figure 6), and the energy consumption of per delivered date byte (Figure 7) in the simulation.

As can be seen from Figure 5, in all system topologies, the average SINR of the system after convergence is obviously higher than that of the initial state owing to the adaptive selection of strategies. In addition, the higher the user density is, the more obvious the effect of our scheme is. In all topologies, the average SINR of the system in the final state is almost 90 dB.

In Figure 6 we compare the obtained average PDR of the system after equilibrium using our scheme against the initial average PDR in different typologies. The result shows that the system after convergence achieves higher average PDR than the system without operating our scheme in all system topologies.

As illustrated in Figure 7, the average energy consumption per delivered data byte of the system in the final state is lower than that of the system that has not performed our scheme. The reason is that the retransmission overhead in the system is reduced by ensuring the QoS of the system after performing our scheme. Moreover, trade-off between obtaining high SINR and maintaining low power utilization also can reduce power dissipation.

5.3. Discussions. From the above described simulation results, we can conclude that, in the system with DISG, an equilibrium state is reached after a few iterations of no-regret learning, where each BSN user selects the suitable strategy to reduce the effect of interuser interference. The system with DISG can provide the appropriate SINR by minimizing the interuser interference, and thus the average PDR is enhanced compared with the system without DISG. The average energy consumption of the system with DISG is reduced owning to the reduction of retransmission overhead as well as the trade-off between SINR and power utilization. In addition, DISG can be applied to different topologies generated by the mobility of BSN users. Therefore, DISG can reduce the effect of interuser interference effectively to ensure specific QoS requirement and save transmission energy at the same time to prolong the lifetime of network.

6. Conclusions and Future Work

We have presented a decentralized interuser interference reduction scheme with noncooperative game for BSN (termed DISG). A no-regret learning algorithm for reducing the effect of the interuser interference with low power consumption has been proposed, and the correctness and effectiveness of DISG have been proved theoretically. The performance of the proposed interuser interference reduction scheme has been evaluated by some experiments performed on the Castalia simulator. Experimental results show that the proposed interuser interference reduction scheme can reduce the effect of the interuser interference effectively by the adaptive selection of strategies to ensure specific QoS requirement, while saving transmission energy and thus prolonging the lifetime of network.

In terms of future work, we would like to implement our scheme on real-life BSNs to further evaluate the performance of the DISG scheme based on more metrics and revise it.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants no. 61173179 and no. 60903153, the Fundamental Research Funds for the Central Universities, and the SRF for ROCS, SEM.

References

[1] P. Kuryloski, A. Giani, R. Giannantonio et al., "Dexternet: an open platform for heterogeneous body sensor networks and its applications," in *Proceedings of the 6th International Workshop on Wearable and Implantable Body Sensor Networks (BSN '09)*, pp. 92–97, IEEE Computer Society, Washington, DC, USA, 2009.

[2] E. Jovanov, C. C. Y. Poon, G. Z. Yang, and Y. T. Zhang, "Guest editorial body sensor networks: from theory to emerging applications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, Article ID 5300655, pp. 859–863, 2009.

[3] D. M. Davenport, B. Deb, and F. J. Ross, "Wireless propagation and coexistence of medical body sensor networks for ambulatory patient monitoring," in *Proceedings of the 6th International Workshop on Wearable and Implantable Body Sensor Networks (BSN '09)*, pp. 41–45, IEEE Computer Society, Washington, DC, USA, 2009.

[4] M. Chen, S. Gonzalez, V. Leung, Q. Zhang, and M. Li, "A 2G-RFID-based e-healthcare system," *IEEE Wireless Communications*, vol. 17, no. 1, Article ID 5416348, pp. 37–43, 2010.

[5] K. Kifayat, P. Fergus, S. Cooper, and M. Merabti, "Body area networks for movement analysis in physiotherapy treatments," in *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA '10)*, pp. 866–872, IEEE Computer Society, Los Alamitos, Calif, USA, 2010.

[6] M. A. Ameen, A. Nessa, and K. S. Kwak, "QoS issues with focus on wireless body area networks," in *Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology (ICCIT '08)*, vol. 1, pp. 801–807, IEEE Computer Society, Washington, DC, USA, 2008.

[7] A. Natarajan, M. Motani, B. de Silva, K. K. Yap, and K. C. Chua, "Investigating network architectures for body sensor networks," in *Proceedings of the 1st ACM SIGMOBILE international Workshop on Systems and Networking Support For Healthcare and Assisted Living Environments (HealthNet '07)*, pp. 19–24, Association for Computing Machinery, San Juan, Puerto Rico, June 2007.

[8] G. Wu, J. Ren, F. Xia, L. Yao, and Z. Xu, "DISG: decentralized inter-user interference suppression in body sensor networks with non-cooperative game," in *Proceedings of the 1st IEEE International Workshop on Mobile Cyber-Physical Systems (MobiCPS '10)*, pp. 256–261, Xi'an, China, October 2010.

[9] S. Razak, V. Kolar, and N. B. Abu-Ghazaleh, "Modeling and analysis of two-flow interactions in wireless networks," *Ad Hoc Networks*, vol. 8, no. 6, pp. 564–581, 2010.

[10] S. Vanka, V. Gupta, and M. Haenggi, "Power-delay analysis of consensus algorithms on wireless networks with interference," *International Journal of Systems, Control and Communications*, vol. 2, no. 1, pp. 256–274, 2010.

[11] H. Boche and M. Schubert, "A unifying approach to interference modeling for wireless networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, Article ID 5433046, pp. 3282–3297, 2010.

[12] S. C. H. Huang, S. Y. Chang, H. C. Wu, and P. J. Wan, "Analysis and design of a novel randomized broadcast algorithm for scalable wireless networks in the interference channels," *IEEE Transactions on Wireless Communications*, vol. 9, no. 7, Article ID 5508973, pp. 2206–2215, 2010.

[13] U. Schilcher, G. Brandner, and C. Bettstetter, "Diversity schemes in interference-limited wireless networks with low-cost radios," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '11)*, pp. 707–712, Cancun, Mexico, March 2011.

[14] G. R. Gupta and N. B. Shroff, "Delay analysis for wireless networks with single hop traffic and general interference constraints," *IEEE/ACM Transactions on Networking*, vol. 18, no. 2, Article ID 5340591, pp. 393–405, 2010.

[15] A. Fasano, "Novel consensus algorithm for wireless sensor networks with noise and interference suppression," in *Proceedings of the IEEE 10th International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '08)*, pp. 416–421, August 2008.

[16] S. Barbarossa and G. Scutari, "Bio-inspired sensor network design," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 26–35, 2007.

[17] L. Liu, X. Zhang, Q. Wang, L. Li, and R. Wang, "CBP-MAC/TFC: a wireless sensor network MAC protocol for systems with burst and periodic signals," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '07)*, pp. 2539–2542, September 2007.

[18] T. Wu and S. Biswas, "Reducing inter-cluster TDMA interference by adaptive MAC allocation in sensor networks," in *Proceedings of the 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*, pp. 507–511, June 2005.

[19] A. K. Sharma, N. Thakral, S. K. Udgata, and A. K. Pujari, "Heuristics for minimizing interference in sensor networks," in *Proceedings of the 10th International Conference on Distributed Computing and Networking (ICDCN '09)*, pp. 49–54, Hyderabad, India, January 2009.

[20] M. K. Marina, S. R. Das, and A. P. Subramanian, "A topology control approach for utilizing multiple channels in multi-radio wireless mesh networks," *Computer Networks*, vol. 54, no. 2, pp. 241–256, 2010.

[21] T. Nguyen, N. Lam, D. Huynh, and J. Bolla, "Minimum edge interference in wireless sensor networks," in *Proceedings of the 5th international conference on Wireless algorithms, Systems, and Applications (WASA '10)*, pp. 57–67, Beijing, China, August 2010.

[22] N. X. Lam, T. N. Nguyen, and D. T. Huynh, "Minimum total node interference in wireless sensor networks," in *Proceedings of the 2nd International ICST Conference on Ad Hoc Networks*, pp. 507–523, Victoria, British Columbia, Canada, August 2010.

[23] O. Begonya, A. Luis, and C. Verikoukis, "Novel qos scheduling and energy-saving mac protocol for body sensor networks optimization," in *Proceedings of the 3rd International Conferenceon Body Area Networks (BodyNets) (ICST '08)*, pp. 1–4, 2008.

[24] B. Otal, L. Alonso, and C. Verikoukis, "Highly reliable energy-saving mac for wireless body sensor networks in healthcare systems," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 4, Article ID 4909290, pp. 553–565, 2009.

[25] B. Braem, B. Latré, C. Blondia, I. Moerman, and P. Demeester, "Improving reliability in multi-hop body sensor networks," in *Proceedings of the 2nd International Conference on Sensor Technologies and Applications (SENSORCOMM '08)*, pp. 342–347, IEEE Computer Society, Washington, DC, USA, 2008.

[26] B. Braem, B. Latre, C. Blondia, I. Moerman, and P. Demeester, "Analyzing and improving reliability in multi-hop body sensor networks," *Advances in Internet Technology*, vol. 2, no. 1, pp. 152–161, 2009.

[27] G. Zhou, J. Lu, C. Y. Wan, M. D. Yarvis, and J. A. Stankovic, "BodyQoS: adaptive and radio-agnostic QoS for body sensor networks," in *Proceedings of the 27th Conference on Computer Communications (INFOCOM '08)*, pp. 565–573, IEEE, 2008.

[28] G. Wu, J. Ren, F. Xia, and Z. Xu, "An adaptive fault-tolerant communication scheme for body sensor networks," *Sensors*, vol. 10, no. 11, pp. 9590–9608, 2010.

[29] J. Kim and Y. Kwon, "Interference-aware topology control for low rate wireless personal area networks," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 1, pp. 97–104, 2009.

[30] D. Djenouri and I. Balasingham, "New QoS and geographical routing in wireless biomedical sensor networks," in *Proceedings of the 6th International Conference on Broadband Communications, Networks and Systems (BROADNETS '09)*, pp. 1–8, Madrid, Spain, 2009.

[31] A. Razzaque, C. S. Hong, and S. Lee, "Data-centric multiobjective QoS-aware routing protocol for body sensor networks," *Sensors*, vol. 11, no. 1, pp. 917–937, 2011.

[32] B. de Silva, A. Natarajan, and M. Motani, "Inter-user interference in body sensor networks: preliminary investigation and an infrastructure-based solution," in *Proceedings of the 6th International Workshop on Wearable and Implantable Body Sensor Networks (BSN '09)*, pp. 35–40, June 2009.

[33] C. A. Otto, E. Jovanov, and A. Milenkovic, "A WBAN-based system for health monitoring at home," in *Proceedings of the 3rd IEEE-EMBS International Summer School and Symposium on Medical Devices and Biosensors (ISSS-MDBS '06)*, pp. 20–23, September 2006.

[34] S. Koskie and Z. Gajic, "A nash game algorithm for SIR-based power control in 3G wireless CDMA networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 5, pp. 1017–1026, 2005.

[35] C. K. Tan, M. L. Sim, and T. C. Chuah, "Game theoretic approach for channel assignment and power control with no-internal-regret learning in wireless ad hoc networks," *IET Communications*, vol. 2, no. 9, pp. 1159–1169, 2008.

[36] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, 1995.

[37] G. Stoltz and G. Lugosi, "Learning correlated equilibria in games with compact sets of strategies," *Games and Economic Behavior*, vol. 59, no. 1, pp. 187–208, 2007.

[38] S. Hart and A. Mas-Colell, "A reinforcement procedure leading to correlated equilibrium," in *Economic Essays: A Festschrift for Werner Hildenbrand*, pp. 181–200, 2001.

[39] H. N. Pham, D. Pediaditakis, and A. Boulis, "From simulation to real deployments in WSN and back," in *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WOWMOM '07)*, pp. 1–6, June 2007.

[40] A. Varga, "The OMNeT++ discrete event simulation system," in *Proceedings of the European Simulation Multicon ference (ESM '01)*, pp. 319–324, 2001.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 980953, 17 pages
doi:10.1155/2011/980953

Research Article

Low-Complexity, Distributed Characterization of Interferers in Wireless Networks

Vibhav Kapnadak,[1] Murat Senel,[2] and Edward J. Coyle[3]

[1] *Network Systems Engineering, AT&T Labs, 2600 Camino Ramon, CA 94583, USA*
[2] *Robert Bosch LLC, Research and Technology Center North America, 4009 Miranda Avenue, Palo Alto, CA 94304, USA*
[3] *School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Drive NW, Atlanta, GA 30332-0250, USA*

Correspondence should be addressed to Edward J. Coyle, ejc@gatech.edu

Received 8 February 2011; Revised 21 May 2011; Accepted 25 May 2011

We consider a large-scale wireless network that uses sensors along its edge to estimate the characteristics of interference from neighboring networks or devices. Each sensor makes a noisy measurement of the received signal strength (RSS) from an interferer, compares its measurement to a threshold, and then transmits the resulting bit to a cluster head (CH) over a noisy communication channel. The CH computes the maximum likelihood estimate (MLE) of the distance to the interferer using these noise-corrupted bits. We propose and justify a low-complexity threshold design technique in which the sensors use nonidentical thresholds to generate their bits. This produces a dithering effect that provides better performance than previous techniques that use different non-identical thresholds or the case in which all the sensor motes use an identical non-optimal threshold. Our proposed technique is also shown (a) to be of low complexity compared to previous non-identical threshold approaches and (b) to provide performance that is very close to that obtained when all sensors use the identical, but unknown, optimal threshold. We derive the Cramér-Rao bound (CRB) and also show that the MLE using our dithered thresholds is asymptotically both efficient and consistent. Simulations are used to verify these theoretical results.

1. Introduction

Large-scale deployments of wireless LANs in unlicensed RF bands are often subject to interference from many sources. We have encountered this problem in the e-Stadium wireless testbed [1], which enables football fans in the stadium at Purdue to access multimedia content related to the game [2] via 802.11b and 3G networks. The locations and coverage areas of the 802.11b access points (APs) in the stadium are shown in Figure 1. The two interfering APs also shown in the figure may disrupt eStadium services for fans sitting along the outer edge of the stands or in the tailgating area north of the stadium. The locations and settings of these interfering access points or devices change over time. The eStadium testbed must sense these changes and adapt its channel assignments and power levels to ensure that its users experience satisfactory Quality of Service (QoS).

The eStadium testbed includes clusters of wireless sensors that are distributed along the concourse area of the stadium [3]. They currently gather and process information from this area and make it available to fans and security personnel during games. We propose that these sensors perform the additional task of characterizing the sources of interference with the 802.11b-based portion of the eStadium testbed.

The information gathered about interferers by these sensors is not directly available to the eStadium APs because of their directional antennas or shadowing by the stadium's structure. A Smart WiFi system [4] would thus not be able to sense and adjust to the presence of these external interferers. This proposed use of the sensors, combined with algorithms to alter the 802.11b channel assignments and power settings, is thus a cognitive networking approach to enabling the coexistence of many systems in unlicensed bands.

The contributions of this paper include the following.

(a) Analysis and improvement of MLE approaches in which sensors conserve energy by minimizing the amount of data they transmit. Each sensor thresholds its noisy measurement of the interferer's signal strength, adds the resulting bit to a packet carrying

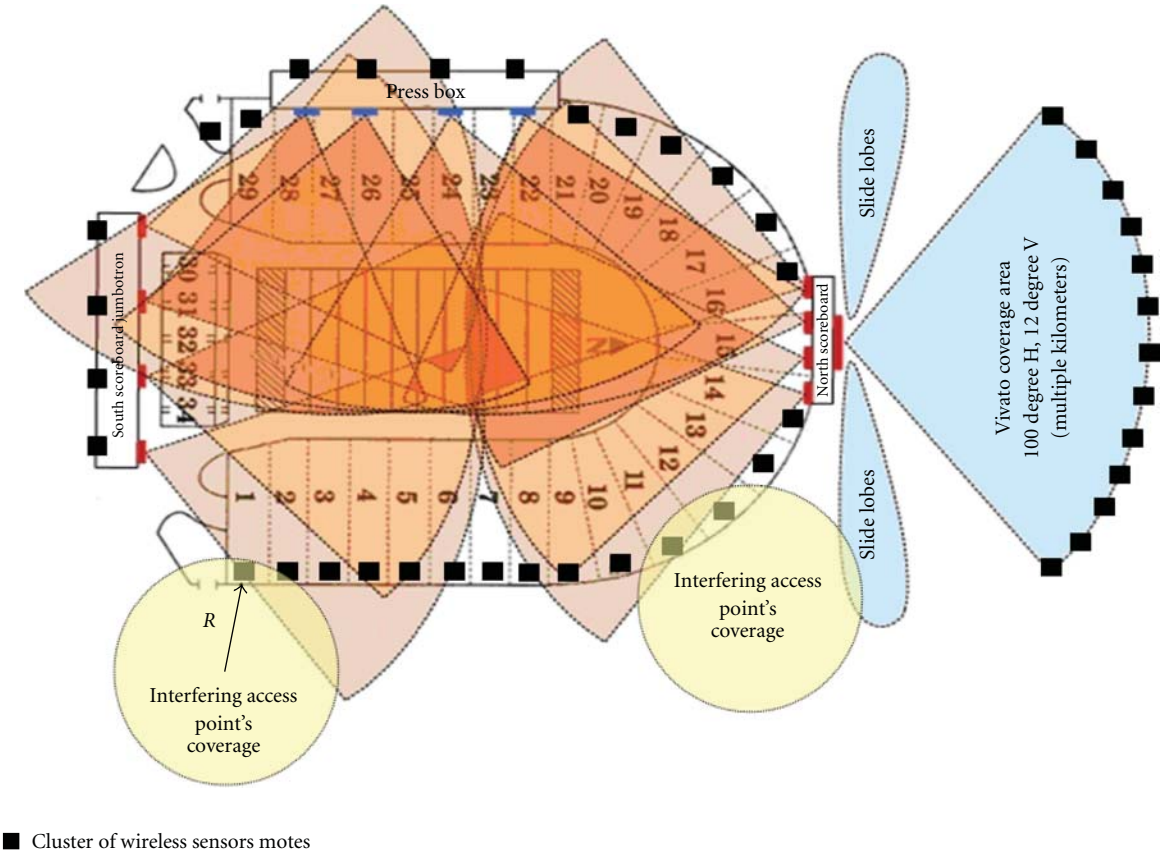


Figure 1: The eStadium testbed at Purdue's Ross Ade stadium includes (1) twelve 802.11 APs with directional antennas that provide coverage, shaded in orange, of the outdoor seating area (2) ten APs with omni antennas that cover the indoor seating areas, (3) a Vivato network that serves the parking area north of the stadium, and (4) clusters of sensors in the concourse area of the stadium. Interfering APs are generally WiFi networks operating in the vicinity of the stadium.

other data, and transmits the packet to the CH via a noisy channel. We consider and justify the case in which each sensor uses a different threshold and compare it with cases in which every sensor uses the same threshold.

(b) Proofs of the asymptotic efficiency and consistency of this new MLE approach and derivation of the Cramér-Rao bound (CRB). Given that the thresholds in our case are non-identical, the standard proofs of asymptotic consistency and efficiency do not hold.

(c) Verification that the MLE based on the use of non-identical thresholds by the sensors performs: (i) as well as one based on the use of identical *optimal* thresholds, even when the number of sensors is small, (ii) significantly better than the one based on the use of identical non-optimal thresholds, and (iii) is of lower complexity and has a performance that is better than or as good as existing non-identical threshold techniques. Because the identical optimal threshold case requires prior knowledge of the true value of the unknown parameter, and existing non-identical techniques involve the solution of a complex optimization problem, our non-identical threshold approach is much more practical and of lower complexity in terms of the design of the thresholds.

In Section 2, we review the most relevant prior work in this area. Sections 3 and 4 describe the system model and the formulation of the sensor fusion problem. The performance of the algorithm is evaluated and compared with existing techniques in Section 5, and conclusions are provided in Section 7.

2. Related Work

In the context of transmitter location estimation for wireless networks, the authors in [5–7] propose expectation maximization (EM) algorithms to locate multiple transmitters using received power levels at arbitrarily placed receivers. The authors in [8] consider a cognitive wireless network in which secondary users measure signal strengths from primary users to estimate the distance to the primary users. They propose estimation algorithms under different channel fading conditions. In [9], the authors study an 802.11 cognitive mesh network (COMNET) in which mesh clients (MC) calculate their distances from the primary stations with triangulation-based localization techniques that use RSS measurements. These localization techniques assume

that the MCs are synchronized and that they are located on the edge of the secondary 802.11 network; otherwise, they may not be able to detect the interference from the primary stations. In the above papers, the authors do not consider the use of thresholded measurements or the transmission of measurements over noisy channels.

In the context of distributed estimation in wireless sensor networks, the authors in [10, 11] consider 1-bit quantization using identical and non-identical thresholds and obtain the maximum-likelihood estimate of an unknown parameter. They account for measurement noise that is Gaussian or has an unknown distribution but do not consider other sources of error, such as noise in the communication channel. A similar situation is examined in [12], where multiple-bit quantized information is used to locate a target in a sensor field under error-free channel conditions. A nonparametric approach to density estimation using a sensor network was introduced in [13]. The authors considered the case where 1-bit quantized data is transmitted to the CH using random thresholds from a pmf. As with the previous papers, however, processing by sensors and the wireless channel between the sensors and the CH are considered to be error free, which is typically not the case in realistic situations. In [14], the authors study channel aware target localization using 1-bit quantized data. They analyze the effects on the root mean squared error (RMSE) of the MLE due to communication channel impairments. They consider the BSC, Rayleigh fading with coherent reception and non-coherent reception. However, in their work, they do not address the design of the 1-bit quantization algorithm for transmission over noisy communication channels.

In [15, 16], the authors consider a 1-bit quantization framework for noisy Gaussian communication links. However, it is assumed that either all sensors use the *same* threshold or use one of two thresholds. More recently, in [17], the authors have proposed an MLE-based distributed estimation algorithm using 1-bit quantized data that is transmitted over a BSC. The context in which the estimation algorithm is analyzed is for secure data transmission in wireless sensor networks, but it is assumed that all the sensors use *identical* thresholds.

Several other papers, including [18–21], also study the distributed estimation problem, but their approach is based on the best linear unbiased estimate (BLUE) at the CH instead of the MLE. This approach is typically not appropriate or optimal when the received data at the CH is a nonlinear function of the unknown parameter.

Our approach is unique because it (a) exploits and justifies the use of different thresholds by the sensors when quantizing their noisy measurements and (b) accounts for noisy measurements by the sensors and error-prone processing and communications between the sensors and the CH. Our analysis of the MLE in this new approach will show that uniformly-spaced thresholds produces a dithering effect that leads to near-optimal performance when the only information available about the parameter being estimated is its support. This is very important because previous approaches based on the use of an identical threshold by all sensors work well *only* when the chosen threshold is either very near the unknown true value, or the number of sensors is very large. Furthermore, our non-identical threshold design algorithm is significantly less complex and performs as good as or better than existing non-identical threshold techniques, such as the one in [10].

3. System Model and Problem Motivation

3.1. System Characteristics and Considerations. The following characteristics of the wireless sensor network and sensor fusion algorithm are assumed in this paper. They are motivated by the eStadium project but are relevant to many other systems operating in unlicensed bands.

(a) *Architecture.* The system consists of a network of wireless APs and a single-hop cluster of N spatially distributed wireless sensors deployed along the edge of the wireless network. Each sensor measures the RSS from an interfering AP, compares it to its threshold, and relays the resulting bit to the CH. The CH fuses the bits it receives to produce an estimate of the signal strength and distance to the interfering AP.

(b) *Measurement Errors.* Due to large-scale fading and shadowing effects, the RSS $X(k)$ observed at sensor k is assumed to be log normally distributed, as in [22, 23]. The mean $\mu(R)$ of this distribution is a function of the distance R from the transmitter to the receiver while the variance σ^2 is independent of R. We assume that $\mu(R)$ follows a path-loss-exponent model; hence,

$$\mu(R) = K - 10\alpha \log(R), \tag{1}$$

where α is the propagation law exponent, and K is the close-in reference power, that is, the power very close to the transmitter. In rest of the paper, the symbol μ will be used interchangeably with $\mu(R)$. On a dB scale, the received RSS at the sensors can be expressed as

$$X_k = \mu(R) + n_k, \quad k = 1 \cdots N, \tag{2}$$

where $\mu(R)$ is the mean and $n_k \sim \mathcal{N}(0, \sigma_n^2)$ is i.i.d. Gaussian noise that is independent of μ. The term σ_n is commonly referred to as the dB spread of the log-normal shadowing and is typically between 4 dB and 12 dB. We refer to this noise as the measurement noise because it corrupts the SNR measurement of each sensor, and we assume that its characteristics are known by the CH. Furthermore, we assume that the sensors are close enough to each other relative to the distance from the interferer that the received RSS at each sensor has the same mean; however, relative to each other, they are spaced far enough to ensure that the noise processes affecting different sensors are independent. The case of dependent noise is left for future work.

(c) *Energy Limits.* Energy efficiency is critical in the design of battery-powered sensor networks, so sensor k thresholds the RSS value it observes to produce only one bit, denoted by b_k, that is to be sent to the CH. The b_k's are transmitted over

a parallel access, noisy communication channel to the CH. These bits may be transmitted individually over the wireless channel or may be combined with other data into packets transmitted over an 802.15.4 (ZigBee) channel [24].

(d) *Communication Channel.* The wireless channel between each sensor, and the CH is modeled as a binary symmetric Channel (BSC) with a crossover probability of ϵ [25]. We assume that the value of ϵ is perfectly estimated at the CH; this can be achieved by the use of pilot training data transmitted by the sensors to the CH during the initialization process. We further assume that the communication channel is static for a period of T seconds; hence, ϵ does not change during the estimation process. For completeness, in Section 5, we also analyze the performance degradation of our algorithms due to a mismatch between the estimated value of ϵ and the true value of ϵ. The choice of a symmetric channel is not required for our analysis but does simplify comparisons. Collisions of bits/packets in the communication channel are assumed to result in their loss and subsequent retransmission. The BSC model thus applies to transmitted bits/packets that are not involved in collisions. Alternatively, one could assume that the sensors' transmissions are scheduled via a collision-free protocol, such as the one in [26], or that they use the TDMA-based frame that is available under the 802.15.4 (ZigBee) protocol [24].

(e) *Reliability and Trustworthiness.* Noisy communications may not be the only source of errors affecting the sensors' decisions before they reach the CH. The sensors may make processing or storage errors. They may be compromised and intentionally report incorrect results with some probability in order to compromise performance without being detected. These additional error sources can be aggregated with the communication errors, resulting in a larger crossover probability. These chapter's results can thus be used to characterize the sensitivity of fusion algorithms to these error sources.

(f) *Industrial, Scientific, and Medical (ISM) Band.* The true mean value of the RSS distribution (see (b), above), denoted by μ_0, is an unknown parameter that lies in the interval $[\mu_l, \mu_u]$. The lower bound, μ_l, corresponds to the maximum range at which the sensors are able to receive and decode a packet from the interferer. The upper bound, μ_u, is related to the maximum transmission power defined by the incumbent network's communication standard. In the eStadium network, these bounds are determined by a typical 802.11b AP operating in the 2.4 GHz band. For a specific example, we use the specs of a Cisco Aironet 1200 series AP [27] with an omni-directional antenna. Its maximum transmission power of 100 mW at a distance of $R = 1$ m from the transmitter corresponds to the upper bound; that is, $\mu_u = -10$ dB or 20 dBm. For the lower bound, the maximum distance at which a typical commercial 802.11 device can receive and decode a packet from an Aironet AP is $R = 200$ m, which corresponds to $\mu_l = -67$ dB or -37 dBm. Figure 2(b) shows the relationship between μ and R for the signal received from a typical AP.

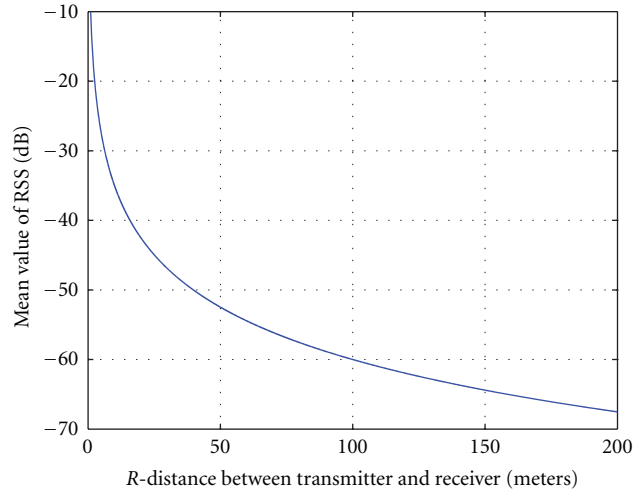


Figure 2: Plot shows the relationship between μ and R with $K = -10$ dB, and $\alpha = 2.5$.

(g) *Periodic Updates.* We assume that the estimates μ and R are updated every T seconds to capture any changes in the set of interferers. T is assumed to be greater than the time required for all sensors to collect a measurement and transmit their bits to the CH. In this paper, we focus on the performance of estimates for a single time period of length T. The better the estimate produced in this period, the faster an iterative algorithm based on it will converge. An iterative algorithm that builds on the results is presented in [28].

In summary, sensor k transmits a single bit of information $b(k)$

$$b_k = \begin{cases} +1 & \text{if } X_k > \tau_k \\ -1 & \text{otherwise,} \end{cases} \tag{3}$$

where τ_k is the threshold used at sensor k to quantize the RSS X_k to one bit. The set of $b(k)'s$ are transmitted to the CH, which uses them to estimate μ and then estimate R via (1). Figure 3 shows the system model and the signal processing block diagram of the cluster of sensors.

3.2. Optimal Identical Threshold Design. The optimal threshold can be defined to be the threshold that minimizes the Cramér-Rao bound (CRB) for any unbiased estimator of μ. This optimization criteria is similar to that proposed in [10, 11] for the error-free communication scenario. In this case, all the sensor motes use the same optimal threshold derived by solving the optimization problem. The thresholds can be expressed as $\tau_k = \tau^{\text{opt}}$ for $k = 1 \cdots N$, where τ^{opt} denotes the optimal threshold. The CRB for any unbiased estimator μ, denoted by $\overline{\mu}$ and derived in Section 5, is given by

$$\text{Var}(\overline{\mu}) \geq \mathcal{J}(\boldsymbol{\tau}, \mu_0)^{-1}, \tag{4}$$

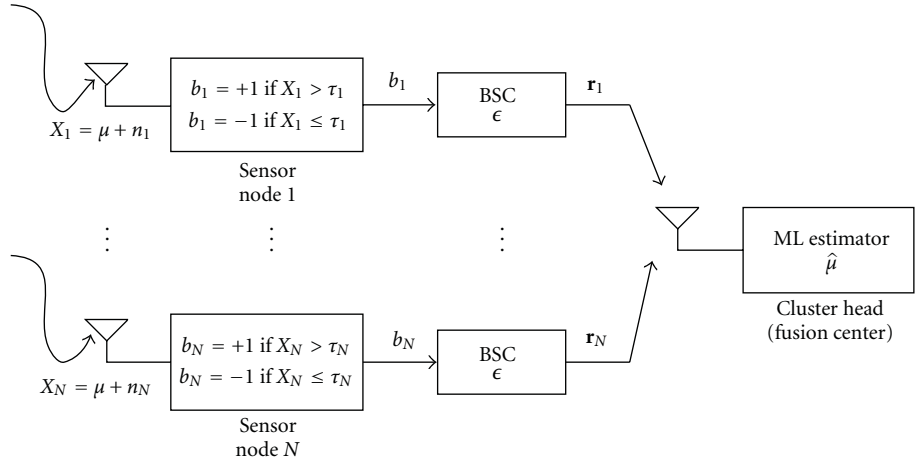


Figure 3: System block diagram for a cluster of wireless sensors. The sensor' RSS measurements are quantized using the dithered quantization technique. The resulting 1-bit values are transmitted over a BSC to the CH, which calculates the MLE of R.

where $\mathcal{I}(\boldsymbol{\tau}, \mu_0)$ is called the Fisher information (FI). As derived later in the paper, the FI is expressed as

$$\mathcal{I}(\boldsymbol{\tau}, \mu_0) = \sum_{k=1}^{N} \mathcal{I}(\tau_k, \mu_0), \tag{5}$$

where

$$\mathcal{I}(\tau_k, \mu_0) = \frac{\left((1-2\epsilon) f_n(\tau_k - \mu_0)\right)^2}{\epsilon + (1-2\epsilon)\overline{F}_n(\tau_k - \mu_0)\left(1 - \epsilon - (1-2\epsilon)\overline{F}_n(\tau_k - \mu_0)\right)} \tag{6}$$

is the FI for sensor k, $\boldsymbol{\tau} = [\tau_1 \cdots \tau_N]$, μ_0 is the true value of μ, and $\overline{F}_n(\cdot)$ and $f_n(\cdot)$ are the complementary cumulative distribution function (ccdf) and pdf of the measurement noise n, respectively. Due to lack of space, we only outline the general technique to obtain the critical points of the CRB function. Calculating the gradient of the r.h.s of (4) with respect to τ_k for $k = 1 \cdots N$, it can easily be shown that when $\tau_k = \mu_0$ for each k, the N equations are simultaneously equal to zero, and the Hessian of the CRB function is positive definite. Hence, the optimal threshold is $\tau^{\text{opt}} = \mu_0$. Note that this choice of threshold is not feasible because μ_0 is not known in advance.

An interesting observation is that the optimal identical threshold choice does not depend on σ_n or ϵ. It is only a function of the true unknown value μ_0. Furthermore, the performance of the identical threshold scheme is very sensitive to the choice of threshold, as shown in Section 6. Even a slight deviation of the threshold value from the true value of μ can result in significant performance degradation. Hence, our goal is to design a quantization scheme which is independent of the true value of μ and has a performance comparable to the optimal identical threshold case.

3.3. Threshold Design for Dithered Quantization. Because use of the optimal identical threshold for the 1-bit quantization step requires prior knowledge of the unknown parameter, we propose a framework in which each sensor uses a different threshold. This ensures that at least a few thresholds will be close to the true value of the unknown parameter. Furthermore, using non-identical thresholds at the sensors for the 1-bit quantization produces a dithering effect that reduces the bias in the estimator (as shown in Section 5). We thus refer to this technique as dithered quantization. Since it is well known that the uniform distribution has maximum entropy among all continuous distributions with compact support [25], we assume that μ is uniformly distributed over the interval $[\mu_l, \mu_u]$ and hence space the quantization thresholds equally over this interval. This leads to the thresholds being assigned for sensor mote k according to the following equation:

$$\tau_k = \mu_l + \frac{k(\mu_u - \mu_l)}{N+1}. \tag{7}$$

3.4. Binning-Based Nonidentical Threshold Design. Another approach to the use of non-identical threshold was proposed in [10]. They consider designing the threshold vector $\boldsymbol{\tau} = \{\tau_k, k \in \mathbb{Z}\}$ and the associated frequency vector $\boldsymbol{\rho} = \{\rho_k, k \in \mathbb{Z}\}$ to minimize the weighted asymptotic variance. The frequency for threshold τ_k is defined as $\rho_k = N_k/N$, where N_k is the total number of sensors transmitting binary information using threshold τ_k, and N is the total number of sensors in the network. Although the problem studied in [10] was for the error-free communication channel scenario, it can be modified for the BSC case. Hence, for our setup, this optimization problem for a uniform weighting function between $[\mu_l, \mu_u]$ denoted by $\mathcal{W}(\mu) = 1/(\mu_u - \mu_l)$ can be formulated as the following second-order cone program (SOCP) with auxiliary variable t as

$$\begin{aligned} \boldsymbol{\rho}^{\star} &= \underset{(t,\boldsymbol{\rho})}{\operatorname{argmin}}\, t, \\ s.t. &\|\mathbf{s} - \mathbf{P}\boldsymbol{\rho}\| \leq t, \\ &\boldsymbol{\rho} \geq \mathbf{0}, \quad \boldsymbol{\rho}^T \mathbf{1} = 1, \end{aligned} \tag{8}$$

where $\mathbf{s} := [S(\mu_0) \cdots S(\mu_M)]^T$, $\boldsymbol{\rho} := [\rho_1 \cdots \rho_L]^T$, L is the number of thresholds, M is the number of discrete points at which the function $S(\mu)$ is evaluated, the function $S(\mu)$ is given by

$$S(\mu) = \mathcal{K}\mathcal{W}^{1/2}(\mu), \qquad \mathcal{K} = \frac{\int_{-\infty}^{+\infty} ((1-2\epsilon)f_n(u))^2 / \Big(\Big(\epsilon + (1-2\epsilon)\overline{F}_n(u)\Big)\Big(1-\epsilon-(1-2\epsilon)\overline{F}_n(u)\Big)\Big)du}{\int_{-\infty}^{+\infty} \mathcal{W}^{1/2}(\mu)\,d\mu}, \tag{9}$$

and $\mathbf{P}$ is a $M \times L$ matrix with its (i, j)th entry given by

$$[\mathbf{P}]_{ij} = \frac{\Big((1-2\epsilon)f_n\Big(\tau_j - \mu_i\Big)\Big)^2}{\Big(\epsilon + (1-2\epsilon)\overline{F}_n\Big(\tau_j - \mu_i\Big)\Big)\Big(1-\epsilon-(1-2\epsilon)\overline{F}_n\Big(\tau_j - \mu_i\Big)\Big)}. \tag{10}$$

Using numerical techniques to approximate the numerator of $\mathcal{K}$, $S(\mu)$ can be expressed as

$$S(\mu) = \frac{\int_{-\infty}^{+\infty} ((1-2\epsilon)f_n(u))^2 / \Big(\Big(\epsilon + (1-2\epsilon)\overline{F}_n(u)\Big)\Big(1-\epsilon-(1-2\epsilon)\overline{F}_n(u)\Big)\Big)du}{|\mu_u - \mu_l|}. \tag{11}$$

For the rest of the paper, we refer to this technique as the binning-based non-identical threshold approach. In Section 5, we compare the performance of all three threshold design schemes.

3.5. Effect of Network Topology on Threshold Design. Wireless sensor networks change with time because sensors can fail as their batteries die, or new sensors may be added to an existing network. Our threshold design is robust to these changes, since when the number of sensors is large, the thresholds are densely distributed over the interval, the loss of a few sensors will not result in a significant loss of information. When the number of sensors is small, the thresholds are coarsely spread over the interval. A few failures might thus result in a significant decrease in the number of bits received by the CH. The CH can then reassign thresholds by communicating with the sensors over the wireless channel. When a new sensor joins the network, it can request a threshold from the CH, choose one based on its unique MAC address, or the CH can reassign all sensors' thresholds according to the algorithm in Section 3.3. On the other hand, the binning-based non-identical threshold technique requires the solution to the optimization problem (8) each time the thresholds need to be assigned. Hence, compared to our proposed approach, the binning-based approach is of much higher computational complexity and therefore cannot easily adapt to changes in the network topology.

4. Sensor Fusion Algorithm

We use maximum-likelihood estimation (MLE) techniques to estimate μ and then use the invariance property of MLE's [29] obtain an estimate for R. For sensor mote k, we denote $\mathcal{R}_k$ to be the random variable (r.v) for the received bit and r_k to be the value of the r.v. Since the received bits $\mathcal{R}_k$ are independent, nonidentically distributed (i.n.i.d) Bernoulli random variables (rv's),

$$\begin{aligned} \Pr(\mathcal{R}_k = 1 \mid \mu) &= \epsilon + (1-2\epsilon)\overline{F}_n(\tau_k - \mu), \\ \Pr(\mathcal{R}_k = -1 \mid \mu) &= (1-\epsilon) - (1-2\epsilon)\overline{F}_n(\tau_k - \mu), \end{aligned} \tag{12}$$

where $\overline{F}_n(\cdot)$ is the complementary cumulative distribution function (ccdf) of the measurement noise $n \sim \mathcal{N}(0, \sigma_n^2)$. Denoting the likelihood function of each received bit by $f(r_k \mid \mu)$, the likelihood function for the received vector of observations $\mathbf{r} = [r_1, \ldots, r_N]$ becomes

$$\begin{aligned} f(\mathbf{r} \mid \mu) &= \prod_{k=1}^{N} f(r_k \mid \mu) \\ &= \prod_{k=1}^{N} \Pr(\mathcal{R}_k = 1 \mid \mu)^{\mathbb{1}_{\{r_k=1\}}} \Pr(\mathcal{R}_k = -1 \mid \mu)^{\mathbb{1}_{\{r_k=-1\}}}. \end{aligned} \tag{13}$$

Let $\mathcal{H}_k(\mu) = \epsilon + (1-2\epsilon)\overline{F}_n(\tau_k - \mu)$, then the log-likelihood function is given by

$$\begin{aligned} &\ln f(\mathbf{r} \mid \mu) \\ &\quad = \sum_{k=1}^{N} [\mathbb{1}_{\{r_k=1\}} \ln(\mathcal{H}_k(\mu)) + \mathbb{1}_{\{r_k=-1\}} \ln(1 - \mathcal{H}_k(\mu))]. \end{aligned} \tag{14}$$

Thus, the MLE $\hat{\mu}$ is obtained by solving the following constrained maximization problem,

$$\begin{aligned} &\underset{\mu}{\text{argmax}} \quad \sum_{k=1}^{N} \left[\mathbb{1}_{\{r_k=1\}} \ln\left(\mathcal{H}_k(\mu)\right) + 1_{\{r_k=1\}} \ln\left(\mathbb{1} - \mathcal{H}_k(\mu)\right)\right], \\ &\text{s.t.} \quad \mu_l \le \mu \le \mu_u, \end{aligned} \tag{15}$$

and because the function relating μ to R in (1) is one to one, the MLE $\hat{R}$ can be obtained directly from $\hat{\mu}$. Note that we have assumed that the CH has perfect knowledge of the dB spread of the log-normal shadowing, that is, σ_n^2 and the cross-over probability ϵ. The former can usually be obtained offline by conducting field measurements in the area of deployment and then performing regression analysis using the experimental data [30]. As for ϵ, this is typically known a priori depending on the modulation and coding scheme used. For example, for transmission of information over fading channels, the probability of bit error (p_e) is well known when standard digital modulation (BSPK, QPSF, 64-QAM, etc.) schemes are used with error control coding strategies such as Turbo or LDPC codes. Hence, p_e can be used as the value for ϵ.

5. Performance Analysis of Dithered Quantization Approach

This section contains (i) proofs of the asymptotic consistency and efficiency of the MLE for μ and R using the proposed dithered quantization approach (ii) derivation of the Cramér-Rao bound (CRB) for the variance of any unbiased estimator for μ and R, and (iii) Monte Carlo simulations that validate the asymptotic results and compare the performance of the dithered quantization technique with the identical threshold-based quantization approach. Since the constrained ML optimization in (15) is analytically intractable, there is no closed-form expression for the estimates or the distribution of the estimates. Hence, to study the bias of the MLE's, we prove that the estimators are asymptotically strongly consistent and hence asymptotically unbiased. Furthermore, although the $\mathcal{R}(k)$'s are i.n.i.d Bernoulli r.v's, we prove that MLE's are asymptotically normally distributed and efficient. Monte Carlo simulations confirm these results and, as shown later, demonstrate that these desirable asymptotic behavior is effectively achieved with as few as $70 \sim 100$ sensors.

5.1. Asymptotic Properties. For the following results, we define $\hat{\mu}_N$, $\hat{R}_N$ to be the MLE of μ and R, respectively, obtained with N sensors using the dithered quantization technique.

Proposition 1. *The MLE $\hat{\mu}_N$ is unique and asymptotically strongly consistent* $\Pr(\lim_{N\to\infty}\hat{\mu}_N = \mu_0) = 1$, *where μ_0 represents the true value of μ.*

Proof. See Appendix A. □

Proposition 2. *$\hat{\mu}_N$ is asymptotically efficient* $(\hat{\mu}_N - \mu_0) \underset{\mathcal{D}}{\rightarrow} \mathcal{N}(0, 1/(\mathcal{I}(\mu_0)))$, *where μ_0 represents the true value of μ, $\mathcal{I}(\cdot)$ is the Fisher information (FI), and $\underset{\mathcal{D}}{\rightarrow}$ means convergence in distribution.*

Proof. See Appendix B. □

The following two corollaries are immediate consequences of the invariance property of functions of MLE's and the one-to-one mapping between μ and R.

Corollary 1. *The MLE $\hat{R}_N$ is unique and asymptotically strongly consistent* $\Pr(\lim_{N\to\infty}\hat{R}_N = R_0) = 1$, *where R_0 represents the true value of R.*

Corollary 2. *$\hat{R}_N$ is asymptotically efficient* $(\hat{R}_N - R_0) \underset{\mathcal{D}}{\rightarrow} \mathcal{N}(0, 1/(\mathcal{I}(R_0)))$, *where R_0 represents the true value of R.*

5.2. Cramér-Rao Bound. In this subsection, we derive the CRB for any unbiased estimator of μ denoted by $\overline{\mu}$. The MLE $\hat{\mu}$ derived previously may be biased for small N, but since we have proved that it is asymptotically unbiased, the CRB is useful as a lower bound of the variance of $\hat{\mu}$ for medium-to-large numbers of sensor motes.

Theorem 1. *The CRB for any unbiased estimator $\overline{\mu}$ obtained using the dithered quantization technique with the BSC model is*

$$\operatorname{Var}(\overline{\mu}) \ge \left[\sum_{k=1}^{N}\left\{\frac{\left((1-2\epsilon)f_n(\tau_k-\mu)\right)^2}{\left(\epsilon+(1-2\epsilon)\overline{F}_n(\tau_k-\mu)\right)\left(1-\epsilon-(1-2\epsilon)\overline{F}_n(\tau_k-\mu)\right)}\right\}\right]^{-1}. \tag{16}$$

Proof. See Appendix C. □

Using the CRB for $\overline{\mu}$ and observing that (1) is a one-to-one transformation, the CRB for any unbiased estimator of R, denoted by $\overline{R}$, can be derived using the invariance property of MLE's and functions of MLE's. The following lemma relates the variance of $\overline{\mu}$ to the variance of $\overline{R}$.

Lemma 1. *If $R = g(\mu)$, then the variance of any unbiased estimate of R is given by*

$$\operatorname{Var}\left(\overline{R}\right) \ge \frac{\left(\partial g/\partial\mu\right)^2}{\mathcal{I}(\mu_0)}. \tag{17}$$

Proof. A proof of this lemma can be found in [29, Chapter 3, Appendix 3A]. □

Corollary 3. *The CRB for any unbiased estimator for $\overline{R}$ using the dithered quantization technique with the BSC model is*

$$\mathrm{Var}\left(\overline{R}\right) \geq \frac{\left((\ln 10)10^{(K-\mu)/10\alpha}/10\alpha\right)^2}{\mathcal{I}(\mu)}, \tag{18}$$

where

$$\mathcal{I}(\mu) = \sum_{k=1}^{N}\left\{\frac{\left((1-2\epsilon)f_n(\tau_k-\mu)\right)_k^2}{\left(\epsilon+(1-2\epsilon)\overline{F}_n(\tau_k-\mu)\right)\left(1-\epsilon-(1-2\epsilon)\overline{F}_n(\tau_k-\mu)\right)}\right\} \tag{19}$$

is the FI.

Proof. Using Lemma 1 and the relationship between R and μ given by $R = 10^{(K-\mu)/10\alpha}$, the result follows. □

Figures 4(a) and 4(b) show the behavior of the square root of the CRB of $\hat{R}$ when we vary the number of sensors, N, and the crossover probability, ϵ, respectively. The CRB is a monotonically increasing function of ϵ for $0 \leq \epsilon < 0.5$ and, due to its symmetry about $\epsilon = 0.5$, is a monotonically decreasing function for $0.5 < \epsilon \leq 1$. As $\epsilon \to 0.5$ the CRB tends to ∞ because, from (1), the Fisher information converges to 0. From Figure 4(a), we observe that the CRB is a monotonically decreasing function of N, which is expected because more sensors means more observations, thus, lower CRB values. Although MLE's tend to have a significant bias for small sample sizes, we have proved that the MLE for our scheme is asymptotically strongly consistent and efficient. Therefore, the analytic derivations above for the CRB for $\overline{\mu}$ and $\overline{R}$ hold in the asymptotic region. The simulation results in the following subsection show that the asymptotic behavior is achieved even with a small number of sensors.

5.3. Numerical Simulation Results. To evaluate the performance of the dithered quantization scheme and compare it with the identical and the binning-based non-identical threshold approaches, we now conduct numerical Monte Carlo simulations. In these simulations (i) the number of sensors is varied from 10 to 200 and (ii) both a low ϵ of 0.05 and a high ϵ of 0.1 are considered. The value of R_0 is set to 50 m, which corresponds to $\mu_0 \approx -52.5$ dB and K is set to -10 dB, the maximum transmission power allowed for 802.11b. We use the typical value of 2.5 for the path-loss exponent α. The standard deviation of the RSS distribution observed at the sensor motes is set to 4 dB; that is, $\sigma_n = 4$ dB.

The results obtained were averaged over 2000 realizations of the measurement noise process. It can easily be shown (see Appendix D) that the log-likelihood function is *not* a concave function of μ for both non-identical threshold-based approaches. Hence, we implemented an 1-dimensional iterated grid search algorithm [31] to locate an approximate maximal value of the log-likelihood function denoted by $\mathcal{L}$. We employed an equidistance grid of size n points given as:

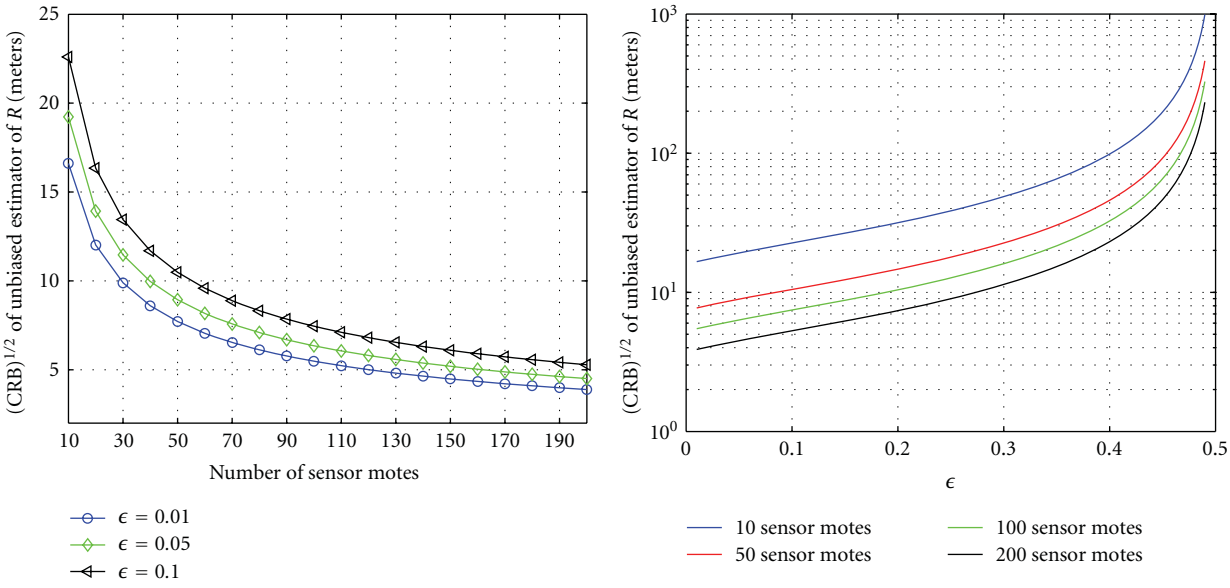
$$\mathcal{S} = \left\{x_k \mid x_k = \mu_l + \frac{k(\mu_u - \mu_l)}{n-1}, \quad k = 0 \cdots n-1\right\}. \tag{20}$$

Using the grid $\mathcal{S}$, the values $\mathcal{L}(x_1), \mathcal{L}(x_2) \cdots \mathcal{L}(x_n)$ are used in the grid search algorithm. An approximate local maximum is located by choosing the grid point x_k that has the largest value for $\mathcal{L}(x_k)$. Using this initialization point, we then used Matlab's sequential quadratic programming (SQP) method to perform the local optimization to determine the MLE.

To characterize the computational complexity of the optimization algorithm, we note that the local optimization using SQP has quadratic convergence rate. The computational complexity of the iterative grid search algorithm can be expressed as $\mathcal{C}(n,r) = rn^m$, where m is the dimension of the parameter space, n is the grid size, and r is the number of iterations. In our simulation experiments, we found that with $r = 10$ and a grid size of $n = 100$ the algorithm converged to an approximate local maximal point quickly. Hence, the computational complexity of the grid search for our case is $O(n)$. In our Matlab implementation of the grid search algorithm, we were able to reduce the complexity to $O(n \log n)$, by the use of sorting algorithms to find the grid point x_k that would correspond to the largest value of the log-likelihood function $\mathcal{L}(x_k)$. In a practical setting, the search grid will be precomputed, and the algorithm will be executed on the CH. Since the CH typically has more processing power than other motes, in a practical setting the convergence of the algorithm would be very fast.

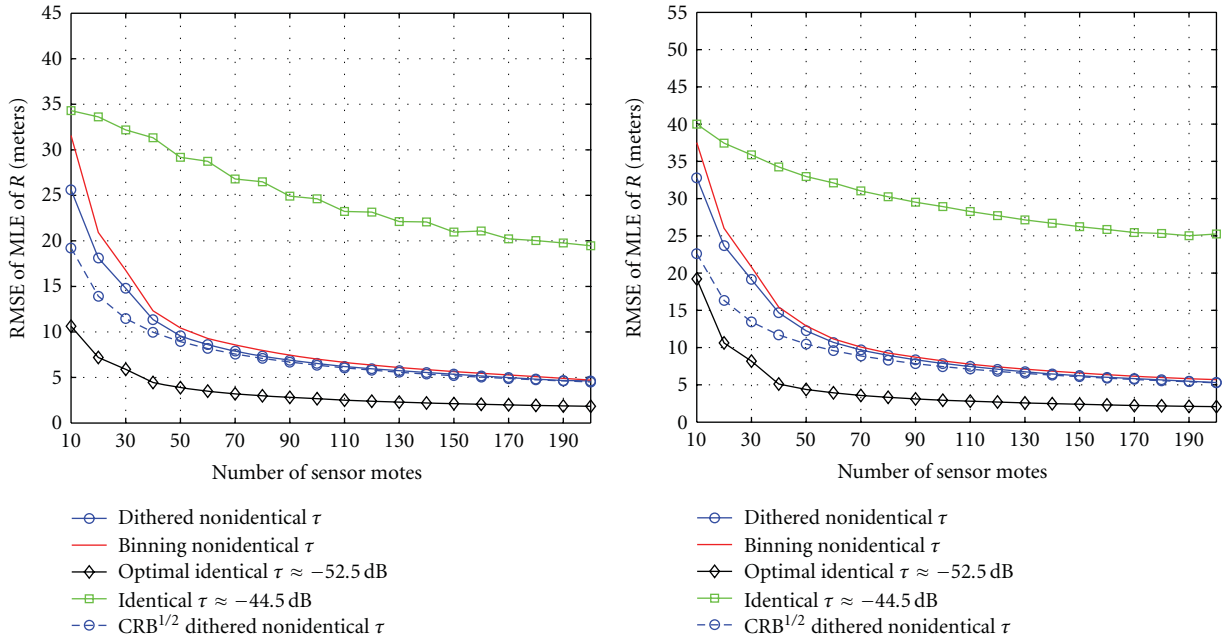
The non-identical thresholds for the dithered quantization technique are generated using the technique described in Section 3, and, for the binning-based technique the SOCP 7 is solved in Matlab using cvx software [32]. As mentioned in [10], the threshold spacing for the binning technique is set to $2\sigma_n$, and hence we have $\tau_{k+1} - \tau_k = 8$ dB, therefore the number of threshold bins $B = (\mu_u - \mu_l)/8 = 8$. For the case where an identical threshold is used by all sensor motes, we consider two scenarios. In the first case, the optimal threshold is used; that is, $\tau = \mu_0$, which corresponds to $\tau \approx -52.5$ dB. In the second case, an identical nonoptimal threshold is used, this corresponds to choosing τ above/below the true mean value of the RSS distribution; that is, $\tau = (\mu_0 + \Delta)$ dB. For simulation purposes, we set $\Delta = 8$; this corresponds to $\tau \approx -44.5$ dB. Similar simulation results are obtained for the case when $\tau > \tau^{\mathrm{opt}}$.

In Figures 5(a) and 5(b), we plot the root mean squared error (RMSE) of $\hat{R}$ for the low and high ϵ cases. We observe that the RMSE using the dithered quantization approach is lower for a small number of sensor motes compared to the non-optimal identical threshold approach and the binning-based non-identical approaches. As the number of sensor motes is increased the RMSE of the binning-based non-identical technique converges to the dithered quantization technique. This is expected since for the binning based approach as the number of sensor motes is increased the threshold frequency (ρ_k) increases which leads to the improvements in the performance.



(a) (b)

FIGURE 4: $(\mathrm{CRB})^{1/2}$ of $\overline{R}$ with $R_0 = 50$ m, $K = -10$ dB, $\alpha = 2.5$, and $\sigma_n = 4$ dB.(a) shows that the analytic expression for the square root of the CRB function is monotonically decreasing function with respect to the number of sensor motes. (b) shows that the analytic expression for the square root of the CRB function is monotonically increasing for the range $0 \leq \epsilon < 0.5$ and as ϵ approaches 0.5 the function approaches ∞ as the Fisher information converges to 0.



(a) (b)

FIGURE 5: RMSE of $\widehat{R}$ with $R_0 = 50$ m, $K = -10$ dB, and $\alpha = 2.5$, $\sigma_n = 4$ dB. In (a) $\epsilon = 0.05$ and in (b) $\epsilon = 0.1$. In both cases the dithered quantization technique has a lower RMSE compared to the non-optimal and binning-based thresholding techniques particularly for a low number of sensor motes.

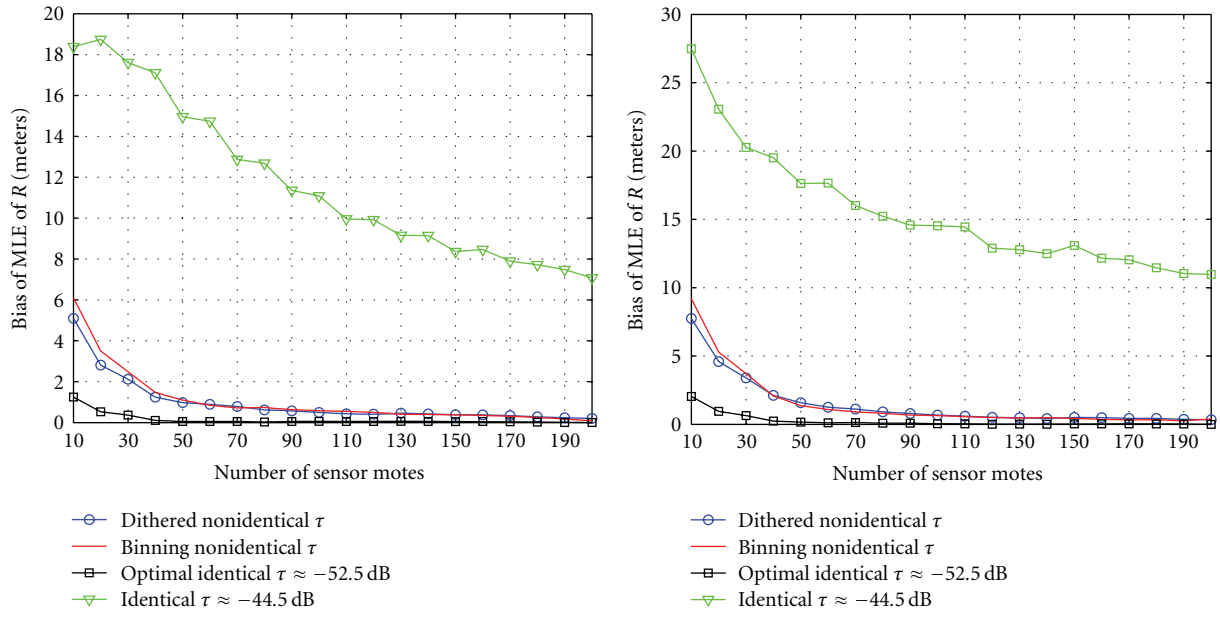


FIGURE 6: Bias of $\hat{R}$ with $R_0 = 50$ m, $K = -10$ dB, and $\alpha = 2.5$, $\sigma_n = 4$ dB. In (a) $\epsilon = 0.05$ and in (b) $\epsilon = 0.1$. In both cases the dithered quantization technique has a lower bias compared to the non-optimal and binning-based thresholding techniques especially for a low number of sensor motes.

Figures 6(a) and 6(b) show the bias of $\hat{R}$ versus the number of sensor motes. We notice that the bias of the MLE using the dithered quantization technique is lower than the non-optimal estimators. Furthermore, we also observe that the bias decreases at an exponential rate with respect to the number of sensor motes and approaches close to 0 with about 100 sensor motes. Figures 5 and 6 also show that the asymptotic properties of the MLE are achievable with about 50 sensor motes for low values of ϵ and about 70 sensor motes for high values of ϵ.

In Figures 7(a) and 7(b), we compare the performance loss incurred using the different thresholding techniques. We define the $\text{CRB}_{\text{Loss}} = \text{CRB}_{\text{non-opt}} - \text{CRB}_{\text{opt}}$, where $\text{CRB}_{\text{non-opt}}$ is the CRB of any unbiased estimator using either the dithered quantization, binning-based or identical non-optimal thresholding techniques, and CRB_{opt} is the CRB of any unbiased estimator using the optimal identical threshold. This metric serves as a good benchmark as, for large number of sensor motes, the MLE's are unbiased, and hence the MSE is equal to the CRB. Figures 7(a) and 7(b) show the CRB_{Loss} versus the number of sensors. We observe that for low and high ϵ values the loss is significantly less for the dithered quantization technique compared with the other schemes, particularly for a small number of sensor motes which is important from a practical point of view.

Figure 8 shows the robustness of our scheme due to a mismatch in the true value of ϵ and the value of ϵ used by the CH for the ML estimation. We notice that even if there is a mismatch, the performance degradation is not significant. The performance of having a perfectly estimated value of ϵ can be replicated using more sensor motes under the condition of a mismatch in ϵ. Furthermore, the gap between the performance between the perfectly estimated ϵ and the mismatched ϵ decreases as the number of sensor motes increases.

In a realistic situation in which the mean of the RSS distribution at the sensor nodes is unknown priori, there is no way to choose the optimal τ for the identical threshold case. It will thus be difficult to achieve good/reliable estimates using non-optimal identical thresholds. In this type of practical scenario, if both approaches (non-optimal identical and non-identical thresholds) use the same number of sensors, and each sensor samples the RSS once, then using the non-identical threshold approach would produce a much more accurate estimate.

Although both non-identical thresholding approaches require fewer sensors than the non-optimal identical threshold technique to achieve near-optimal performance, we observe that compared to the binning approach our thresholding scheme is of low complexity, as it does not involve solving any complex optimization problem to design the thresholds. Furthermore, the dithered quantization technique performs better in terms of having a lower RMSE, variance, and bias compared with the binning and non-optimal identical thresholding techniques, particularly when the number of sensors is small, which is the typical case in a practical scenario.

In summary, the dithered quantization technique has significant advantages over other techniques in terms of either the number of sensors required to achieve a given accuracy or the accuracy of estimates achievable with a fixed number of sensors.

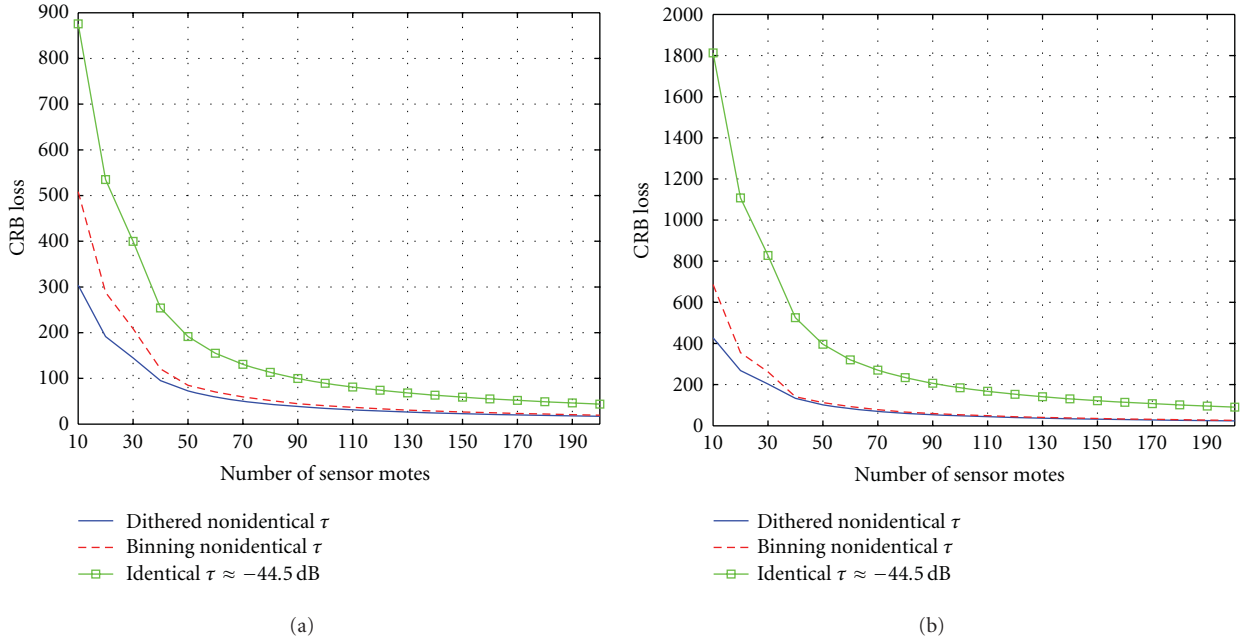


FIGURE 7: $(\text{CRB})^{1/2}$ with $R_0 = 50$ m, $K = -10$ dB, $\alpha = 2.5$, and $\sigma_n = 4$ dB. In (a) $\epsilon = 0.05$ and in (b) $\epsilon = 0.1$, in both cases the loss incurred using the dithered quantization technique is significantly less compared to the other techniques, particularly for low number of sensor motes.

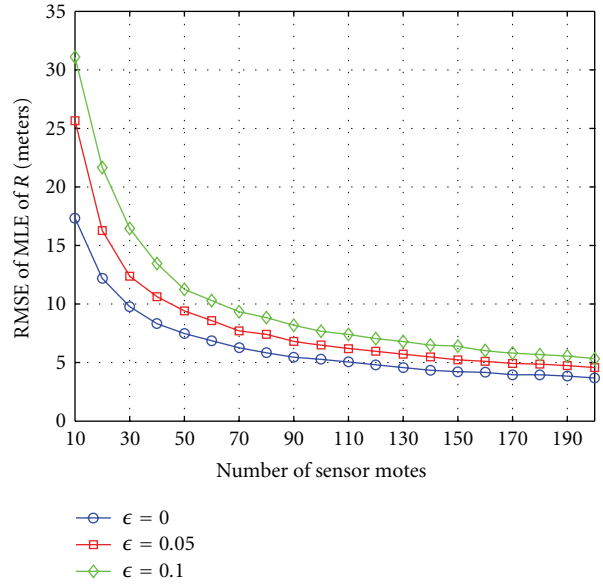


FIGURE 8: shows the RMSE of $\hat{R}$ using dithered quantization technique with $R_0 = 50$ m, $K = -10$ dB, $\alpha = 2.5$, and $\sigma_n = 4$ dB for various values of ϵ. The plot shows the robustness of our estimation scheme due to a mismatch in the true value of ϵ and the estimated value of ϵ used by the CH.

6. Drawbacks of the Identical Threshold Approach

In this section, we study the disadvantages of using the identical threshold approach and the effect on $\hat{R}$ using this technique. As observed in previous work [10, 11], with the error-free communication channel, the disadvantage of the identical threshold approach is that the CRB for $\hat{\mu}$ grows exponentially with $[(\tau - \mu_0)^2/\sigma_n]^2$. A similar observation can be seen for the CRB of $\overline{R}$ when the identical threshold is

used under a BSC model. Using the tight Chernoff bound for the ccdf of the Gaussian distribution (i.e., $\overline{F}_n(\tau-\mu) \leq e^{-(\tau-\mu)^2/2}/2$), the CRB for $\overline{R}$ can be bounded as follows:

$$\begin{aligned}
&\mathrm{CRB}(\epsilon,\tau,R_0)\\
&= G\frac{\left[\epsilon+(1-2\epsilon)\overline{F}_n(\tau-\mu_0)\right]\left[(1-\epsilon)-(1-2\epsilon)\overline{F}_n(\tau-\mu_0)\right]}{N\left[(1-2\epsilon)f_n(\tau-\mu_0)\right]^2}\\
&\leq G\frac{\pi\sigma_n^2 e^{(\tau-\mu_0)^2/2\sigma_n^2}}{2N(1-2\epsilon)^2}\\
&\quad\times\left[\left(2\epsilon+(1-2\epsilon)e^{-(\tau-\mu_0)^2/2\sigma_n^2}\right)\right.\\
&\qquad\left.\times\left(2(1-\epsilon)-(1-2\epsilon)e^{-(\tau-\mu_0)^2/2\sigma_n^2}\right)\right],
\end{aligned} \tag{21}$$

where $G = ((\ln 10)10^{(K-\mu_0)/10\alpha}/10\alpha)^2$. Figure 9(a) shows the exponential increase with respect to $(\tau-\mu_0)/\sigma_n$ for $\epsilon = 0.01$ and $\epsilon = 0.1$ for the CRB of $\hat{R}$. We notice that the Chernoff bound is tight and that a small deviation of τ from the true unknown value μ_0 will result in a significant loss in performance. Although a similar result has previously been widely reported for the error-free channel case [10, 11], to the best of our knowledge, no one has studied the performance of the MLE's in terms of the MSE and variance effects using the identical threshold approach. Hence, to further analyze the effects of using this technique, Figure 9(b) shows the MSE and variance plot as $(\tau-\mu_0)/\sigma_n$ is varied. As with the CRB plot, the performance degradation in terms of the MSE is also exponential as τ deviates from μ_0. However, we notice the existence of multiple critical points. This effect is because the MSE can be decomposed into the $(\text{bias})^2$ and variance. The figure shows that for small deviations from μ_0 the $(\text{bias})^2$ is nearly zero, and the variance increases exponentially; however, when $\tau \gg \mu_0$, the variance starts to decrease, but the bias increases linearly with $(\tau-\mu_0)$. These combined effects reiterate the fact that using identical thresholds could potentially result in estimators with very high bias and MSE.

7. Conclusion

In this paper, we proposed and analyzed a sensor fusion algorithm for interference characterization in wireless networks. To minimize energy usage, each wireless sensor in the network transmits only 1-bit of quantized, noisy RSS information to the FC over a BSC.

Unlike previous approaches that used identical thresholds for all sensors or use non-identical thresholds based on solving a complex optimization problem, we propose a simple and low complexity threshold design technique of uniformly distributing the thresholds over the range of possible values of the mean of the RSS distribution. This approach performs (i) as well as the optimal identical threshold approach and (ii) significantly better than the identical non-optimal threshold design approaches and has a performance that is more accurate or as good as previously proposed non-identical threshold techniques. The optimal identical threshold case is highly unrealistic because the optimal threshold cannot be known in advance; our non-identical threshold technique is much more practical, reliable, and of low complexity design compared to previously proposed techniques based on either using non-identical or identical non-optimal thresholds.

Appendices

A. Asymptotic Consistency of MLE

Proof. We start with the following assumptions that are restated from [33, page 443-444] for completeness and are easy to verify the following.

(A0) The distributions $F(\mathcal{R} \mid \mu)$ of the observations are distinct.

(A1) The distributions $F(\mathcal{R} \mid \mu)$ have common support.

(A2) The observations are $\mathbf{r} = (r_1,\ldots,r_N)$, where the r_k are independent with probability density $f(r_k \mid \mu)$ with respect to the underlying probability measure.

(A3) The parameter space Ω contains an open set ω of which the true parameter value μ_0 is an interior point.

Next, we prove the following lemma.

Lemma 2. *For the independent nonidentically distributed (i.n.i.d) random variables $\mathcal{R}_k$'s we have*

$$\Pr\left\{\lim_{N\to\infty}\ln f\left(\overline{\mathcal{R}} \mid \mu\right) < \ln f\left(\overline{\mathcal{R}} \mid \mu_0\right)\right\} = 1, \quad \forall \mu \neq \mu_0, \tag{A.1}$$

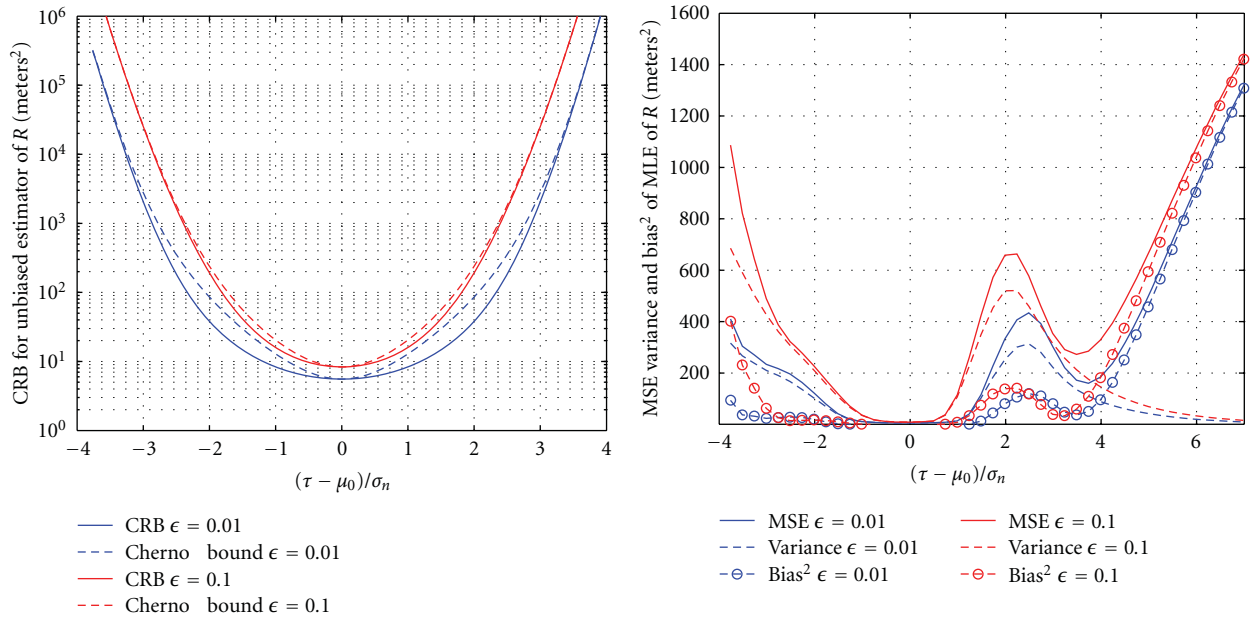
where μ_0 is the true value of μ.

Proof. To prove the lemma, it is equivalent to prove the following:

$$\Pr\left\{\lim_{N\to\infty}\left[\frac{1}{N}\sum_{k=1}^{N}\ln\frac{f(\mathcal{R}_k \mid \mu)}{f(\mathcal{R}_k \mid \mu_0)} < 0\right]\right\} = 1, \quad \forall \mu \neq \mu_0. \tag{A.2}$$

Next, we show that the strong law of large numbers (SLLNs) can be applied to the term $(1/N)\sum_{k=1}^{N}\ln f(\mathcal{R}_k \mid \mu)/f(\mathcal{R}_k \mid \mu_0)$. To show this, we check the Kolmogorov sufficient conditions for the SLLN to hold for independent r.v.'s. Let $Y_k = \ln f(\mathcal{R}_k \mid \mu)/(f(\mathcal{R}_k\mu_0))$, $p_k = \Pr(\mathcal{R}_k = 1 \mid \mu)$, $q_k = \Pr(\mathcal{R}_k = -1 \mid \mu) = (1-p_k)$, $p_{k_0} = \Pr(\mathcal{R}_k = 1 \mid \mu_0)$ and $q_{k_0} = \Pr(\mathcal{R}_k = -1 \mid \mu_0)$. We further assume w.l.o.g. that $\epsilon \leq 0.5$, and using (12) we have the following inequalities:

$$\begin{aligned}
\epsilon \leq p_k \leq 1-\epsilon,\\
1-\epsilon \leq q_k \leq \epsilon.
\end{aligned} \tag{A.3}$$



(a) (b)

FIGURE 9: In (a) the CRB and the Chernoff bound for $\hat{R}$ is shown for varyious values of τ at $R_0 = 50$ m ($\mu_0 \approx -52.5$ dB), $N = 100$ and $\sigma_n = 4$ dB. Observe that the performance degradation is exponential as $(\tau - \mu_0)/\sigma_n$ increases for different crossover probabilities. (b) shows the MSE, variance and (bias)^2 for $\hat{R}$ with varying τ with $R_0 = 50$ m ($\mu_0 \approx -52.5$ dB) $N = 100$ and $\sigma_n = 4$ dB. The MSE also has an exponential growth with respect to $(\tau - \mu_0)/\sigma_n$; however, we notice the peculiar behavior of multiple critical points due to the bias and variance effect of the MLE of R.

Similar inequalities hold for p_{k_0} and q_{k_0}, respectively. The variance of Y_k can be bounded as follows:

$$\begin{aligned}
\mathrm{Var}(Y_k) &\le \mathbb{E}_{\mu_0}\left[{Y_k}^2\right] \\
&= p_{k_0}\left(\ln\frac{p_k}{p_{k_0}}\right)^2 + q_{k_0}\left(\ln\frac{q_k}{q_{k_0}}\right)^2 \\
&\le (1-\epsilon)\left(\ln\frac{(1-\epsilon)}{\epsilon}\right)^2 + \epsilon\left(\ln\frac{\epsilon}{(1-\epsilon)}\right)^2 \\
&= 1.
\end{aligned} \tag{A.4}$$

Therefore, $\lim_{N\to\infty}\sum_{k=1}^{N}\mathrm{Var}(Y_k)/k^2 \le \lim_{N\to\infty}\sum_{k=1}^{N}(1/k^2) < \infty$, hence, the SLLN holds for Y_k's, and we have

$$\Pr\left\{\lim_{N\to\infty}\frac{1}{N}\left[\sum_{k=1}^{N}\ln\frac{f(\mathcal{R}_k \mid \mu)}{f(\mathcal{R}_k \mid \mu_0)} - \sum_{k=1}^{N}\mathbb{E}_{\mu_0}\left(\ln\frac{f(\mathcal{R}_k \mid \mu)}{f(\mathcal{R}_k \mid \mu_0)}\right) = 0\right]\right\} = 1, \tag{A.5}$$

where $\mathbb{E}_{\mu_0}(\cdot)$ represents the expectation using the pdf when $\mu = \mu_0$. Now by Jensen's inequality we have

$$\begin{aligned}
&\mathbb{E}_{\mu_0}\left[\ln\frac{f(\mathcal{R}_k \mid \mu)}{f(\mathcal{R}_k \mid \mu_0)}\right] \\
&< ln\mathbb{E}_{\mu_0}\left[\frac{f(\mathcal{R}_k \mid \mu)}{f\left(\mathcal{R}_k \mid \mu_0\right)}\right] \\
&= \ln\left[\frac{\Pr(\mathcal{R}_1 = 1 \mid \mu)}{\Pr(\mathcal{R}_1 = 1 \mid \mu_0)}\Pr(\mathcal{R}_1 = 1 \mid \mu_0)\right. \\
&\qquad \left. + \frac{\Pr(\mathcal{R}_1 = -1 \mid \mu)}{\Pr(\mathcal{R}_1 = -1 \mid \mu_0)}\Pr(\mathcal{R}_1 = -1 \mid \mu_0)\right] \\
&= \ln[\Pr(\mathcal{R}_1 = 1 \mid \mu) + \Pr(\mathcal{R}_1 = -1 \mid \mu)] \\
&= 0.
\end{aligned} \tag{A.6}$$

Using the result (A.6) in (A.5), statement (A.2) is proved, and hence the lemma follows. □

Next we consider a δ-neighborhood about μ_0 such that $(\mu_0 - \delta, \mu_0 + \delta) \in \omega$ and define $S_N = \{\mathbf{r} : \ln f(\mathbf{r} \mid \mu_0 - \delta) < \ln f(\mathbf{r} \mid \mu_0)$ and $\ln f(\mathbf{r} \mid \mu_0) > \ln f(\mathbf{r} \mid \mu_0 + \delta)\}$. Hence, by Lemma 2 we have,

$$\Pr\left\{\lim_{N\to\infty} S_N \mid \mu_0\right\} = 1 \tag{A.7}$$

and, therefore, for all $\mathbf{r} \in S_N$, there exists $\hat{\mu}_N \in (\mu_0 - \delta, \mu_0 + \delta)$ s.t. the likelihood function is maximized in the interval. Since for any $\delta > 0$ small enough, there exists a sequence $\hat{\mu}_N = \hat{\mu}_N(\delta)$ of roots to the equation $\partial \ln f(\mathbf{r} \mid \mu)/\partial\mu = 0$ s.t. $\Pr\{\lim_{N\to\infty} \mid \hat{\mu}_N - \mu_0 \mid < \delta \mid \mu_0\} = 1$. Next we choose a sequence which does not depend on δ, by observing that the likelihood function is a continuous function, and the limit of a sequence of roots is also another root of the equation $\partial \ln f(\mathbf{r} \mid \mu)/\partial\mu = 0$. Therefore, there exists a root μ_N^* that is closest to μ_0; hence, $\Pr(\lim_{N\to\infty} \mid \mu_N^* - \mu_0 \mid < \delta \mid \mu_0) = 1$ and since δ was arbitrary chosen the proposition follows. □

B. Asymptotic Efficiency of MLE

Proof. The key step is to show that the central limit theorem (CLT) holds for the i.n.i.d log-likelihood functions of the $\mathcal{R}_k$'s. As before, we state the following regularity conditions from [33, page 440-441] for completeness.

(a) The parameter space Ω is an open interval (not necessarily finite).

(b) The distributions of $F(\mathcal{R}_k \mid \mu)$ have common support, so that the set $A = \{r_k : f(r_k \mid \mu) > 0\}$ is independent of μ.

(c) For every $r_k \in A$, the density $f(r_k \mid \mu)$ is twice differentiable with respect to μ, and the second derivative is continuous in μ

(d) $\mathbb{E}[\partial \ln f(\mathcal{R}_k \mid \mu)/\partial\mu] = 0$ and $\mathbb{E}[-\partial^2 \ln f(\mathcal{R}_k \mid \mu)/\partial\mu^2] = \mathbb{E}[(\partial \ln f(\mathcal{R}_k \mid \mu)/\partial\mu)^2] = \mathcal{I}(\mu)$.

(e) The Fisher information $0 < \mathcal{I}(\mu) < \infty$.

(f) For any given $\mu_0 \in \Omega$, there exists a positive number c and a function $M(r_k)$ (both of which may depend on μ_0) s.t.

$$\left| \frac{\partial^2 \ln f(r_k \mid \mu)}{\partial\mu^2} \right| \le M(r_k), \quad \forall r_k \in A,\ \mu_0 - c < \mu < \mu_0 + c, \tag{B.1}$$

and $\mathbb{E}_{\mu_0}[M(\mathcal{R}_k)] < \infty$.

Since most of the conditions listed above are easy to verify, we provide proofs only for (d) and (f). To prove the first part of (d) we need to show that $\mathbb{E}_{\mu_0}[\partial \ln f(\mathcal{R}_k \mid \mu)/\partial\mu] = 0$. Using the definitions for p_k and q_k from Appendix A, we have $p_k' = (1-2\epsilon) f_n(\tau_k - \mu)$, $q_k' = -(1-2\epsilon) f_n(\tau_k - \mu)$, $p_k'' = (1-2\epsilon) f_n(\tau_k - \mu)/\sigma_n^2$, and $q_k'' = p_k''$, where the notation $f' = \partial f/\partial\mu$ and $f'' = \partial^2 f/\partial\mu^2$. Therefore, using (12), we have

$$\begin{aligned}
\mathbb{E}_{\mu_0}\left[\frac{\partial \ln f(\mathcal{R} \mid \mu)}{\partial\mu}\right] &= p_{k_0}\left(\frac{p_k'}{p_k}\right) + q_{k_0}\left(\frac{q_k'}{q_k}\right) \\
&= p_{k_0}\left(\frac{p_k'}{p_k}\right) + (1 - p_{k_0})\left(\frac{p_k'}{(1-p_k)}\right) \\
&= 0.
\end{aligned} \tag{B.2}$$

Similarly we have

$$\begin{aligned}
\mathbb{E}_{\mu_0}\left[\left(\frac{\partial \ln f(\mathcal{R} \mid \mu)}{\partial\mu}\right)^2\right] &= \frac{p_{k_0} {p_k'}^2}{{p_k}^2} + \frac{q_{k_0} {q_k'}^2}{{q_k}^2}, \\
\mathbb{E}\left[\frac{-\partial^2 \ln f(\mathcal{R}_k \mid \mu)}{\partial\mu^2}\right] &= -\left(\frac{p_{k_0}\left(p_k'' p_k - {p_k'}^2\right)}{{p_k}^2} + \frac{q_{k_0}\left(q_k'' q_k - {q_k'}^2\right)}{{q_k}^2}\right) \\
&= -\frac{p_{k_0} p_k''}{p_k} + \frac{q_{k_0} p_k''}{q_k} + \frac{p_{k_0} {p_k'}^2}{{p_k}^2} + \frac{q_{k_0} {q_k'}^2}{{q_k}^2} \\
&= \frac{p_{k_0} {p_k'}^2}{{p_k}^2} + \frac{q_{k_0} {q_k'}^2}{{q_k}^2}.
\end{aligned} \tag{B.3}$$

To check condition (f) using the result from Appendix D we have for all $r_k \in A$, $\mu_0 - c < \mu < \mu_0 + c$

$$\begin{aligned}
\left| \frac{\partial^2 \ln f(r_k \mid \mu)}{\partial\mu^2} \right| &\le \mathbb{1}_{\{r_k = 1\}}\left(\frac{(\mu_u - \mu_l)(1-2\epsilon)}{\sqrt{2\pi}\sigma_n^3}\right) \\
&\quad + \mathbb{1}_{\{r_k = -1\}}\left(\frac{(\mu_u - \mu_l)(1-2\epsilon)}{\sqrt{2\pi}\sigma_n^3}\right).
\end{aligned} \tag{B.4}$$

If we let $M(r_k) = \mathbb{1}_{\{r_k=1\}}((\mu_u - \mu_l)(1-2\epsilon)/\sqrt{2\pi}\sigma_n^3) + \mathbb{1}_{\{r_k=-1\}}((\mu_u - \mu_l)(1-2\epsilon)/\sqrt{2\pi}\sigma_n^3)$, and hence

$$\begin{aligned}
\mathbb{E}_{\mu_0}[M(R_k)] &= p_{k_0}\left(\frac{(\mu_u - \mu_l)(1-2\epsilon)}{\sqrt{2\pi}\sigma_n^3}\right) \\
&\quad + q_{k_0}\left(\frac{(\mu_u - \mu_l)(1-2\epsilon)}{\sqrt{2\pi}\sigma_n^3}\right) < \infty.
\end{aligned} \tag{B.5}$$

Next we prove the CLT for independent nonidentically distributed (i.n.i.d) r.v.'s, which will be used to show the asymptotic efficiency result.

Lemma 3. *For the independent, nonidentically distributed (i.n.i.d) log-likelihood function* $\ln f(\overline{\mathcal{R}} \mid \mu_0)$*, we have*

$$\frac{1}{\sqrt{N}} \frac{\partial}{\partial\mu_0} \ln f\left(\overline{\mathcal{R}} \mid \mu_0\right) \underset{\mathcal{D}}{\to} \mathcal{N}\left(0, \frac{\mathcal{I}(\mu_0)}{N}\right), \tag{B.6}$$

where μ_0 *is the true value of* μ, $\mathcal{I}(\mu_0) = \sum_{k=1}^{N} (1-2\epsilon)^2 f_n^2(\tau_k - \mu_0)/p_{k_0} q_{k_0}$ *is the Fisher information matrix,* $p_{k_0} = \Pr(\mathcal{R}_k = 1 \mid \mu_0)$, $q_{k_0} = \Pr(\mathcal{R}_k = -1 \mid \mu_0)$, *and* $\underset{\mathcal{D}}{\to}$ *means convergence in distribution.*

Proof. For ease of notation let $Y_k = \partial/\partial\mu_0 \ln f(\mathcal{R}_k \mid \mu_0)$, $\sigma_{Y_k}^2 = \mathbb{E}_{\mu_0}[Y_k^2]$, due to the regularity conditions, we have $\mathbb{E}_{\mu_0}[Y_k] = 0$. Therefore, the l.h.s of (B.6) can be written as

$$\frac{1}{\sqrt{N}} \frac{\partial}{\partial\mu_0} \ln f\left(\overline{\mathcal{R}} \mid \mu_0\right) = \frac{\sqrt{N}}{N} \sum_{k=1}^{N} Y_k. \tag{B.7}$$

Since the Y_k's are not i.i.d., to apply the CLT we must verify the Lindeberg sufficiency condition. Hence we need to show that for all $\epsilon > 0$,

$$L(N) = \frac{1}{S_N^2}\sum_{k=1}^{N}\mathbb{E}_{\mu_0}\left\{|Y_k|^2\right\}\mathbb{1}_{\{|Y_k|>\epsilon S_N\}} \underset{N\to\infty}{\to} 0, \tag{B.8}$$

where $S_N^2 = \sum_{k=1}^{N}\sigma_{Y_k}^2$. It can be shown that $\sigma_{Y_k}^2 = (1-2\epsilon)^2 f_n^2(\tau_k-\mu_0)/p_{k_0}q_{k_0}$, where $p_{k_0} = \Pr(\mathcal{R}_k = 1 \mid \mu_0)$ and $q_{k_0} = \Pr(\mathcal{R}_k = -1 \mid \mu_0)$. Using this we have

$$S_N^2 = \sum_{k=1}^{N}\frac{(1-2\epsilon)f_n^2(\tau_k-\mu_0)}{p_{k_0}q_{k_0}} \overset{(a)}{\geq} 4(1-2\epsilon)\sum_{k=1}^{N} f_n^2(\tau_k-\mu_0), \tag{B.9}$$

where (a) in (B.9) follows from the fact that $1/p_{k_0}q_{k_0} \geq 4$ for $p_{k_0}, q_{k_0} \in (0,1)$. Therefore,

$$\begin{aligned} L(N) &= \frac{1}{S_N^2}\sum_{k=1}^{N}\int_{\{|Y_k|>\epsilon S_N\}}|Y_k|^2 dF(Y_k) \\ &= \frac{1}{S_N^2}\sum_{k=1}^{N}\int_{\mathbb{R}}|Y_k|^2\mathbb{1}_{\{|Y_k|>\epsilon S_N\}}dF(Y_k). \end{aligned} \tag{B.10}$$

We observe that from (B.9) that $S_N^2 \underset{N\to\infty}{\to} \infty$ hence

$$\lim_{N\to\infty} L(N) = \lim_{N\to\infty}\frac{1}{S_N^2}\sum_{k=1}^{N}\int_{\mathbb{R}}|Y_k|^2\mathbb{1}_{\{|Y_k|>\epsilon S_N\}}dF(Y_k) = 0, \tag{B.11}$$

where (B.11) follows from the fact that $\lim_{N\to\infty}\int_{\mathbb{R}}|Y_k|^2\mathbb{1}_{\{|Y_k|>\epsilon S_N\}}dF(Y_k) = 0$ by the Lebesgue dominated convergence theorem. Hence, the Lindeberg conditions are satisfied and the result follows. □

For rest of the proof, let $\mathcal{L}'(\mu) = \sum_{k=1}^{N}(\partial/\partial\mu)\ln f(\mathcal{R}_k \mid \mu)$, $\mathcal{L}''(\mu) = \sum_{k=1}^{N}(\partial^2/\partial\mu^2)\ln f(\mathcal{R}_k \mid \mu)$. Then, by the mean value theorem, $\exists\ \lambda \in (0,1)$ with $\mu_N^* = \mu_0 + \lambda(\hat{\mu}_N - \mu_0)$ for some $\mu_N^* \in (\mu_0, \hat{\mu}_N)$ such that $\mathcal{L}''(\hat{\mu}_N)$:

$$\mathcal{L}'(\hat{\mu}_N) = \mathcal{L}'(\mu_0) + (\hat{\mu}_N - \mu_0)\mathcal{L}''(\mu_N^*). \tag{B.12}$$

Since $\hat{\mu}_N$ is the MLE, the l.h.s. of (B.12) is equal to 0. We thus obtain

$$\sqrt{N}(\hat{\mu}_N - \mu_0) = \frac{(1/\sqrt{N})\mathcal{L}'(\mu_0)}{-(1/N)\mathcal{L}''(\mu_N^*)}. \tag{B.13}$$

Now since $\hat{\mu}_N$ is a strongly consistent estimator of μ_0, we have $\mu_N^* \underset{\mathcal{P}}{\to} \mu_0$, where $\underset{\mathcal{P}}{\to}$ means convergence in probability. This implies $-(1/N)\mathcal{L}''(\mu_N^*) \underset{\mathcal{P}}{\to} -(1/N)\mathcal{L}''(\mu_0)$, by the fact that if $\mu_N^* \underset{\mathcal{P}}{\to} \mu_0$, then $g(\mu_N^*) \underset{\mathcal{P}}{\to} g(\mu_0)$ for a continuous function $g(\cdot)$. Using arguments similar to those in Appendix A, it can be shown that the Kolmogorov sufficient conditions for the SLLN are satisfied for $(1/N)\mathcal{L}''(\mu_0)$, hence $-(1/N)\mathcal{L}''(\mu_0) \underset{\mathcal{P}}{\to} \mathcal{I}(\mu_0)/N$. By the CLT proved earlier, we have $(1/\sqrt{N})\mathcal{L}'(\mu_0) \underset{\mathcal{D}}{\to} \mathcal{N}(0, \mathcal{I}(\mu_0)/N)$. Hence by Slutsky's theorem, we thus have $\sqrt{N}(\hat{\mu}_N - \mu_0) \underset{\mathcal{D}}{\to} \mathcal{N}(0, N/\mathcal{I}(\mu_0))$ and the proposition follows. □

C. Cramér-Rao Bound

Proof. Using the definition of Fisher information matrix, we have

$$\begin{aligned} -\mathbb{E}_{\mu}\left[\frac{\partial^2 \ln f\left(\overline{\mathcal{R}} \mid \mu\right)}{\partial\mu^2}\right] &= -\sum_{k=1}^{N}\mathbb{E}_{\mu_0}\left[\frac{\partial^2 \ln f(\mathcal{R}_k \mid \mu)}{\partial\mu^2}\right] \\ &= -\sum_{k=1}^{N}\left[\frac{\partial^2 \ln f(\mathcal{R}_k = 1 \mid \mu)}{\partial\mu^2}\Pr(\mathcal{R}_k = 1 \mid \mu)\right. \\ &\qquad \left. + \frac{\partial^2 \ln f(\mathcal{R}_k = -1 \mid \mu)}{\partial\mu^2}\Pr(\mathcal{R}_k = -1 \mid \mu)\right]. \end{aligned} \tag{C.1}$$

Differentiating $\ln f(\mathcal{R}_k = 1 \mid \mu)$ and $\ln f(\mathcal{R}_k = -1 \mid \mu)$ with respect to μ twice, we have:

$$\begin{aligned} \frac{\partial^2 \ln f(\mathcal{R}_k = 1 \mid \mu)}{\partial\mu^2} &= \frac{-(1-2\epsilon)^2 f_n^2(\tau_k-\mu) + \Pr(\mathcal{R}_k = 1 \mid \mu)(1-2\epsilon)f_n'(\tau_k-\mu)}{P(\mathcal{R}_k = 1 \mid \mu)^2}, \\ \frac{\partial^2 \ln f(\mathcal{R}_k = -1 \mid \mu)}{\partial\mu^2} &= \frac{-(1-2\epsilon)^2 f_n'(\tau_k-\mu)\Pr(\mathcal{R}_k = -1 \mid \mu) - (1-2\epsilon)^2 {f_n}^2(\tau_k-\mu)}{\Pr(\mathcal{R}_k = -1 \mid \mu)^2}. \end{aligned} \tag{C.2}$$

Using (C.2) and (C.1) can be expressed as

$$\mathcal{I}(\mu_0) = -\sum_{k=1}^{N}\left\{\frac{((1-2\epsilon)f_n(\tau_k-\mu))^2}{\left(\epsilon+(1-2\epsilon)\overline{F}_n(\tau_k-\mu)\right)\left(1-\epsilon-(1-2\epsilon)\overline{F}_n(\tau_k-\mu)\right)}\right\}. \tag{C.3}$$

Hence Theorem 1 is proved. □

D. Concavity of Log-Likelihood Function

Lemma 4. *The log-likelihood function* $\ln f(\mathbf{r} \mid \mu)$ *is not a concave function.*

Proof. Let $G(\mu) = \sum_{k=1}^{N}(\mathbb{1}_{\{r_k=1\}}\ln\alpha_k(\mu) + \mathbb{1}_{\{r_k=-1\}}\ln\beta_k(\mu))$, where $\alpha(\mu) = \epsilon + (1-2\epsilon)\overline{F}_n(\tau_k-\mu)$, $\beta(\mu) = (1-\epsilon) - (1-2\epsilon)\overline{F}_n(\tau_k-\mu)$, and w.l.o.g we assume that $\epsilon < 0.5$. From the properties of concave functions, we know that if $\alpha(\mu)$ and $\beta(\mu)$ are concave and positive, then $G(\mu)$ is concave. Consider the terms $\alpha_k(\mu)$ and $\beta_k(\mu)$, the first and second derivatives of these terms can be expressed as

$$\begin{aligned}
\alpha_k'(\mu) &= (1-2\epsilon)f_n(\tau_k-\mu),\\
\alpha_k''(\mu) &= \frac{(\tau_k-\mu)\alpha_k'(\mu)}{\sigma_n^2},\\
\beta_k'(\mu) &= -(1-2\epsilon)f_n(\tau_k-\mu) = -\alpha_k'(\mu),\\
\beta_k''(\mu) &= \frac{-(\tau_k-\mu)\alpha_k'(\mu)}{\sigma_n^2}.
\end{aligned} \tag{D.1}$$

Using the above equations, we can express the second derivative of the log-likelihood function as

$$\begin{aligned}
G''(\mu) = \sum_{k=1}^{N}\Big(&\mathbb{1}_{\{r_k=1\}}\left[A_k(\mu)(\tau_k-\mu) - (B_k(\mu))^2\right]\\
&+\mathbb{1}_{\{r_k=-1\}}\left[-C_k(\mu)(\tau_k-\mu) - (D_k(\mu))^2\right]\Big),
\end{aligned} \tag{D.2}$$

where $A_k(\mu) = \alpha_k'(\mu)/\sigma_n^2\alpha_k(\mu)$, $B_k(\mu) = \alpha_k'(\mu)/\alpha_k(\mu)$, $C_k(\mu) = \alpha_k'(\mu)/\sigma_n^2\beta_k(\mu)$, and $D_k(\mu) = \alpha_k'(\mu)/\beta_k(\mu)$. For $G(\mu)$ to be concave, a sufficient condition is that for all $\mu \in [\mu_l, \mu_u]$, $G''(\mu) \le 0$, it can be seen from (D.2) that depending on the values of τ_k and the received bits r_k the sufficiency condition may not be satisfied. For example, consider the scenario with $N = 2$ in this case $\tau_1 \approx -29.2$ dB, and $\tau_2 \approx -48.4$ dB. With the received vector $\mathbf{r} = [r_1 r_2] = [1\,1]$, the second derivative of the log-likelihood function can be expressed as

$$\begin{aligned}
G''(\mu) &= \frac{\alpha_1'(\mu)(\tau_1-\mu)}{\sigma_n^2\alpha_1(\mu)} - \left(\frac{\alpha_1'(\mu)}{\alpha_1(\mu)}\right)^2\\
&\quad+\frac{\alpha_2'(\mu)(\tau_2-\mu)}{\sigma_n^2(1-\alpha_2(\mu))} - \left(\frac{\alpha_2'(\mu)}{(1-\alpha_2(\mu))}\right)^2.
\end{aligned} \tag{D.3}$$

An equivalent sufficiency condition to show that $G(\mu)$ is concave is if $-G''(\mu) \ge 0$, for all $\mu \in [\mu_l, \mu_u]$. Using the definitions of $\alpha_k(\mu)$ and $\alpha_k'(\mu)$, we can obtain the following upper bounds

$$\begin{aligned}
\alpha_k'(\mu) &\le (1-2\epsilon)f_n(0) = \frac{(1-2\epsilon)}{\sqrt{2\pi}\sigma_n},\\
\frac{1}{(1-\epsilon)} &\le \frac{1}{\alpha_k(\mu)} \le \frac{1}{\epsilon}.
\end{aligned} \tag{D.4}$$

Then $-G''(\mu)$ can be upper bounded as follows:

$$\begin{aligned}
-G''(\mu) &= \frac{\alpha_1'(\mu)(\mu-\tau_1)}{\sigma_n^2\alpha_1(\mu)} + \left(\frac{\alpha_1'(\mu)}{\alpha_1(\mu)}\right)^2 + \frac{\alpha_2'(\mu)(\mu-\tau_2)}{\sigma_n^2(1-\alpha_2(\mu))}\\
&\quad+\left(\frac{\alpha_2'(\mu)}{(1-\alpha_2(\mu))}\right)\\
&\le \frac{(1-2\epsilon)(\mu-\tau_1)}{\epsilon\sqrt{2\pi}\sigma_n^3} + \left(\frac{1-2\epsilon}{\sqrt{2\pi}\sigma_n\epsilon}\right)^2\\
&\quad+\frac{(1-2\epsilon)(\mu-\tau_2)}{\epsilon\sqrt{2\pi}\sigma_n^3} + \left(\frac{1-2\epsilon}{\sqrt{2\pi}\sigma_n\epsilon}\right)^2\\
&= \frac{(1-2\epsilon)}{\epsilon\sigma_n^2\sqrt{\pi}}\left(\frac{2\mu-(\tau_1+\tau_2)}{\sqrt{2}\sigma_n} + \frac{(1-2\epsilon)}{\epsilon\sqrt{\pi}}\right)\\
&\approx \frac{(1-2\epsilon)}{\epsilon\sigma_n^2\sqrt{\pi}}\left(\frac{2\mu+77.6}{\sqrt{2}\sigma_n} + \frac{(1-2\epsilon)}{\epsilon\sqrt{\pi}}\right).
\end{aligned} \tag{D.5}$$

It can be seen from (D.5) that for $\mu_l \le \mu \le \sqrt{2}\sigma_n(2\epsilon-1)/(2\epsilon\sqrt{\pi}) - 38.8$ the second derivative of the log-likelihood function $-G'' < 0$ and hence $G(\mu)$ will not be concave for this range of μ. With $\epsilon = 0.1$, we have that for $\mu_l \le \mu < -51.6$ dB $G''(\mu) > 0$. Therefore, we can conclude that log-likelihood function is not concave for all $\mu \in [\mu_l, \mu_u]$. □

References

[1] X. Zhong, H. H. Chan, T. J. Rogers, C. P. Rosenberg, and E. J. Coyle, "The development and eStadium testbeds for research and development of wireless services for large-scale sports venues," in *Proceedings of the 2nd International IEEE/Create-Net Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM '06)*, pp. 340–348, March 2006.

[2] A. Ault, J. V. Krogmeier, S. R. Dunlop, and E. J. Coyle, "eStadium: the mobile wireless football experience," in *Proceedings of the 3rd International Conference on Internet and Web Applications and Services (ICIW '08)*, pp. 644–649, June 2008.

[3] X. Zhong and E. J. Coyle, "eStadium: a wireless "living lab" for safety and infotainment applications," in *Proceedings of the 1st International Conference on Communications and Networking in China (ChinaCom '06)*, 2007.
[4] A. Hills, "Smart Wi-Fi," *Scientific American*, vol. 285, no. 10, pp. 86–94, 2005.
[5] J. K. Nelson and M. R. Gupta, "An EM technique for multiple transmitter localization," in *Proceedings of the 41st Annual Conference on Information Sciences and Systems (CISS '07)*, pp. 610–615, March 2007.
[6] J. K. Nelson, M. U. Hazen, and M. R. Gupta, "Global optimization for multiple transmitter localization," in *Proceedings of the Military Communications Conference (MILCOM '06)*, pp. 1–7, October 2006.
[7] T. Roos, P. Myllymaki, and H. Tirri, "A statistical modeling approach to location estimation," *IEEE Transactions on Mobile Computing*, vol. 1, no. 1, pp. 59–69, 2002.
[8] X. Liu and S. Shankar, "Sensing-based opportunistic channel access," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 577–591, 2006.
[9] K. R. Chowdhury and I. F. Akyildiz, "Cognitive wireless mesh networks with dynamic spectrum access," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 168–181, 2008.
[10] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks—part I: gaussian case," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1131–1143, 2006.
[11] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks—part II: unknown probability density function," *IEEE Transactions on Signal Processing*, vol. 54, no. 7, pp. 2784–2796, 2006.
[12] R. Niu and P. K. Varshney, "Target location estimation in sensor networks with quantized data," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4519–4528, 2006.
[13] A. Dogandžić and B. Zhang, "Nonparametric probability density estimation for sensor networks using quantized measurements," in *Proceedings of the 41st Annual Conference on Information Sciences and Systems (CISS '07)*, pp. 759–764, March 2007.
[14] O. Ozdemir, R. Niu, and P. K. Varshney, "Channel aware target localization with quantized data in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1190–1202, 2009.
[15] T. C. Aysal and K. E. Barner, "Constrained decentralized estimation over noisy channels for sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1398–1410, 2008.
[16] T. C. Aysal and K. E. Barner, "Blind decentralized estimation for bandwidth constrained wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, Article ID 4524301, pp. 1466–1471, 2008.
[17] T. C. Aysal and K. E. Barner, "Sensor data cryptography in wireless sensor networks," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, Article ID 4470151, pp. 273–289, 2008.
[18] Z. Q. Luo, "Universal decentralized estimation in a bandwidth constrained sensor network," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2210–2219, 2005.
[19] S. Cui, J. J. Xiao, A. J. Goldsmith, Z. Q. Luo, and H. V. Poor, "Estimation diversity and energy efficiency in distributed sensing," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4683–4695, 2007.
[20] J. J. Xiao, S. Cui, Z. Q. Luo, and A. J. Goldsmith, "Power scheduling of universal decentralized estimation in sensor networks," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 413–422, 2006.
[21] J. J. Xiao and Z. Q. Luo, "Decentralized estimation in an inhomogeneous sensing environment," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3564–3575, 2005.
[22] E. Visotsky, S. Kuffher, and R. Peterson, "On collaborative detection of TV transmissions in support of dynamic spectrum sharing," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 338–345, 2005.
[23] A. Ghasemi and E. S. Sousa, "Asymptotic performance of collaborative spectrum sensing under correlated log-normal shadowing," *IEEE Communications Letters*, vol. 11, no. 1, pp. 34–36, 2007.
[24] IEEE802.15.4, "Ieee standard 802 part 15.4 wireless medium access control (mac) and physical layer (phy) specifications for low rate wireless personal area networks (wpans)," 2006.
[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.
[26] S. Bandyopadhyay, Q. Tian, and E. J. Coyle, "Spatio-temporal sampling rates and enery efficiency in wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, pp. 1339–1352, 2005.
[27] Cisco, "Cisco aironet 1200 series access point data sheet," [Online], 2006, http://www.cisco.com/en/US/prod/collateral/wireless/ps5678/ps430/ps4076/product_data_sheet09186a00800937a6.html.
[28] V. Kapnadak, M. Senel, and E. J. Coyle, "Distributed iterative quantization for interference characterization in wireless networks," *Digital Signal Processing*. In press.
[29] S. M. Kay, *Fundamentals of Statistical Signal Processing-Estimation Theory*, vol. 1, PrenticeHall PTR, Upper Saddle River, NJ, USA, 1993.
[30] T. S. Rappaport, *Wireless Communications- Principles and Practices*, PrenticeHall PTR, Upper Saddle River, NJ, USA, 2002.
[31] R. A. Thisted, *Elements of Statistical Computing*, vol. 1, Chapman and Hall, 1988.
[32] "Cvx: Matlab software for disciplined convex programming," http://cvxr.com/cvx/.
[33] E. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, 1998.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 946258, 8 pages
doi:10.1155/2011/946258

Research Article

Modeling and Performance Analysis of Energy Efficiency Binary Power Control in MIMO-OFDM Wireless Communication Systems

Yuming Wang,[1] Xiaohu Ge,[1] Chengqian Cao,[1] Xi Huang,[1] and Frank Y. Li[2]

[1] *Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*
[2] *Department of Information and Communication Technology, University of Agder (UiA), 4879 Grimstad, Norway*

Correspondence should be addressed to Xiaohu Ge, xhge@mail.hust.edu.cn

Received 31 March 2011; Accepted 30 June 2011

The energy efficiency optimization of the binary power control scheme for MIMO-OFDM wireless communication systems is formulated, and then a global optimization solution of power allocation is derived. Furthermore, a new energy efficiency binary power control (EEBPC) algorithm is designed to improve the energy efficiency of MIMO-OFDM wireless communication systems. Simulation results show that the EEBPC algorithm has better energy efficiency and spectrum efficiency than the average power control algorithm in MIMO-OFDM wireless communication systems.

1. Introduction

It is beyond question that information and communication technology (ICT) industries play a significant role in current global economy. Among all energy-consuming industries, the ICT industry takes 2% of global total CO2 emissions while consuming 3% of global energy storage [1, 2]. Within the 3% consumption, 57% is caused in mobile and wireless communication systems [3]. From another perspective of ICT growth, there has risen a high demand for broadband data transmission with high-quality services, which is further triggering the Multi-Input and Multi-Output (MIMO) and Orthogonal Frequency Division Multiplexing (OFDM) techniques to be adopted in the next generation wireless communication systems [4–6]. Therefore, the optimization of energy efficiency in MIMO-OFDM wireless communication systems has become an urgent requirement.

In wireless communication systems, the transmission power consumption of base station is controlled by power allocation schemes. In early analog mobile communication systems, the key aim of transmission power allocation scheme is to improve user signal-to-noise (SNR). Therefore, some transmission power allocation schemes of base station are developed to enhance the SNR of terminal users [7–9]. In digital mobile communication systems, the traditional transmission power allocation schemes sought to realize the maximization of wireless channel capacity [10–12]. One of the most popular power allocation schemes of base station is the power water-filling scheme, which performs power allocation based on the state of wireless channels to close the wireless channel capacity to the Shannon capacity limit [10]. To overcome the requirement of continuous-rate adaptation in the power water-filling scheme, a new scheme based on a fixed number of codes, was proposed to maximize average spectral efficiency (ASE) of dual-branch MIMO systems with perfect transmitter and receiver channel state information (CSI) [11]. To formulate the link adaptation problem as a convex optimization problem, Kim and Daneshrad proposed a link adaptation power strategy to maximize energy efficiency or data throughput subject to a given quality of service (QoS) constraint [12]. Considering the complexity of optimization transmission power allocation scheme, a simple transmission power allocation scheme, that is, the binary power control scheme was proposed to maximize wireless channel capacity in practical engineering applications [13, 14]. However, the energy efficiency problem in the traditional binary power control scheme of MIMO-OFDM wireless communication systems is not considered.

In this paper, we investigate the energy efficiency problem of the binary power control scheme and formulate the binary

power control scheme with energy efficiency constraint. Moreover, a new algorithm is designed to address the energy efficiency power allocation in MIMO-OFDM wireless communication systems. The contributions and novelty of this paper are summarized as follows.

(1) We formulate the energy efficiency problem of the binary power control scheme in MIMO-OFDM wireless communication systems.

(2) A global optimization solution of power allocation in MIMO-OFDM wireless communication systems is derived. Furthermore, two derivation results can be used for potential engineering application with the low calculation complexity.

(3) A new algorithm is designed to realize the energy efficiency binary power control scheme in communication systems.

(4) Performance of the new algorithm is analyzed and some interesting observations are presented.

The rest of paper is organized as follows. In Section 2, the energy efficiency concept is introduced and the binary power control scheme is introduced. In Section 3, we investigate optimal conditions for energy efficiency transmission with the binary power control scheme. Moreover, the global optimization solution of power distribution model is derived and a new algorithm is proposed. Furthermore, we apply the new algorithm in a MIMO-OFDM communication system and provide simulation results to demonstrate energy efficiency improvement in Section 4. Finally, we conclude the paper in Section 5.

2. Energy Efficiency in Wireless Communication Systems

As introduced in this section, more and more energy efficiency optimization schemes in wireless communication systems were studied. To estimate the energy efficiency, the definition of energy efficiency should be first declared. In this paper, a definition of energy efficiency is described as follows.

2.1. Definition of Energy Efficiency in Wireless Communication Systems. Typically, the energy consumption of transmitting per bit is a main concern in evaluating the energy efficiency of a communication system. In addition, the definition should include the transmission power from the base station and the capacity of wireless channels. Considering the Shannon capacity theory [15], the maximum achievable capacity of a wireless channel is related to the transmission power from a base station. Therefore, an energy efficiency used in wireless communication systems is defined as follows:

$$\eta = f(P_i) = \frac{\sum_{i=1}^{n} C_i(P_i)}{\sum_{i=1}^{n} P_i}, \quad i \in [1, n], \tag{1}$$

where η is the energy efficiency of wireless communication systems which is denoted as a function of transmission power P_i over wireless channel $i \cdot n$ is the number of wireless channels in wireless communication systems. $C_i(P_i)$ is a wireless channel capacity which is denoted as a function of transmission power P_i over wireless channel i.

2.2. Binary Power Control Scheme. For wireless communication systems, the transmission power P over wireless channels is allocated from the minimum value $P_{\min}$ to the maximum value $P_{\max}$, that is, $P_{\min} \le P \le P_{\max}$. According to the binary power control scheme of wireless communication systems, the transmission power P over wireless channels is allocated a value from two elements, that is, $P_{\min}$ and $P_{\max}$. Moreover, the binary power control scheme is formulated as follows [16]:

$$P \in \Omega, \quad \Omega = \{P \mid P = P_{\min} \text{ or } P = P_{\max}\}. \tag{2}$$

From research results in [13, 14, 17], when there are many wireless channels in wireless communication systems, the capacity and rate performance of wireless communication systems adopting the binary power control scheme can approximate the Shannon capacity limit. However, when the energy efficiency of wireless communication systems is considered as an optimal aim, how to adopt the binary power control scheme of wireless communication systems to approximate the global energy efficiency optimal solution is a great challenge.

3. Problem Formulation of Energy Efficiency Binary Power Control Scheme

To investigate the binary power control scheme in energy efficiency of wireless communication systems, a single cell MIMO-OFDM wireless communication system is illustrated in Figure 1. One base station integrated with M_T antennas is located in the center of cell. There are K users uniformly scattering in the cell and every user is integrated with M_R antennas. To simplify the modeling complexity of the OFDM scheme, all orthogonal N subcarriers are regrouped into N subchannels by the OFDM scheme. Moreover, interference between users is assumed to be ignored in this single cell. Every subchannel of MIMO-OFDM communication system in Figure 1 is assumed as a quasistatic channel, which means there is no change within a block of transmission. The bandwidth of MIMO-OFDM communication system is normalized as 1. For one moment, without loss of generality, only N subchannels are enabled for data transmission. The CSI of MIMO-OFDM wireless communication system is assumed to be known by the base station in Figure 1. In this paper, our research focuses on the downlink performance of wireless communication systems.

3.1. Problem Formulation. Based on the system model in Figure 1, the total capacity of MIMO-OFDM communication system is described as

$$C_{\text{total}} = \sum_{i=1}^{N} \log_2\left(1 + \frac{P_i}{n_0} \|H_i\|_F^2\right), \tag{3}$$

where P_i is the transmission power over wireless subchannel i, n_0 is the additive white Gaussian noise (AWGN) in

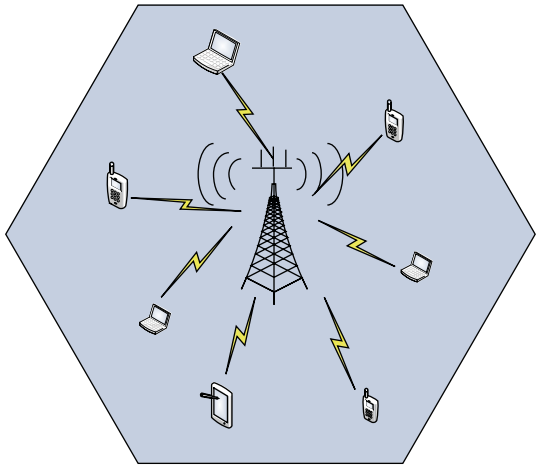


Figure 1: System model of MIMO-OFDM wireless communication systems.

wireless subchannels, and $\|H_i\|_F$ is the F norm over wireless subchannel i.

In this case, the total transmission power in the downlink of MIMO-OFDM communication system is denoted as

$$P_{\text{total}} = \sum_{i=1}^{N} P_i. \tag{4}$$

Furthermore, the energy efficiency of MIMO-OFDM communication system is given by

$$\eta_{\text{total}} = \frac{\sum_{i=1}^{N} \log_2\left(1 + (P_i/n_0)\|H_i\|_F^2\right)}{\sum_{i=1}^{N} P_i}. \tag{5}$$

The wireless subchannel set K is defined as follows:

$$\text{CH}_i \in K, \quad K = \left\{\text{CH} \mid \text{CH} = \bigcup_{i=1}^{N} \text{CH}_i\right\}, \tag{6}$$

where CH is the wireless subchannel element in set K and CH_i is the wireless subchannel i.

Assume that the binary power control scheme is used to allocate the transmission power for the wireless subchannel set K. In this case, the set K is divided into two subsets: one is the maximum power transmission subchannel subset $K_{p\max}^{M}$ with M wireless subchannels; the other is the minimum power transmission subchannel subset $K_{p\min}^{N-M}$ with $N - M$ wireless subchannels. Moreover, the total transmission power of $K_{p\max}^{M}$ is denoted as $P_{\max\_\text{total}}$ and the total transmission power of $K_{p\min}^{N-M}$ is denoted as $P_{\min\_\text{total}}$. The relationship of P_{total}, $P_{\max\_\text{total}}$, and $P_{\min\_\text{total}}$ is described as follow:

$$\begin{aligned} P_{\text{total}} &= P_{\max\_\text{total}} + P_{\min\_\text{total}}, \\ P_{\max\_\text{total}} &= M \times P_{\max}, \\ P_{\min\_\text{total}} &= (N - M) \times P_{\min}. \end{aligned} \tag{7}$$

To optimize the global energy efficiency of MIMO-OFDM wireless communication system, some basic assumptions and a principle are defined as follows.

Assumption 1. The total transmission power of MIMO-OFDM wireless communication system is fixed as a constant.

Assumption 2. $P_{\max\_\text{total}}$ in the maximum power transmission subchannel subset is fixed as a constant.

Principle 1. A wireless subchannel CH_k is assigned into the maximum power transmission subchannel subset $K_{p\max}^{M}$ only when the energy efficiency of $K_{p\max}^{M}$ including CH_k is no less than the energy efficiency of $K_{p\max}^{M}$ without CH_k, otherwise, the wireless subchannel should be assigned into the minimum power transmission subchannel $K_{p\min}^{N-M}$.

3.2. Optimization of Power Allocation. For the energy efficiency binary power control scheme, how to optimize the power allocation in the maximum power transmission subchannel subset and the minimum power transmission subchannel subset is a key problem. Based on assumption, principle and (7) in Section 3.1, we find that $P_{\max\_\text{total}}$ and $P_{\min\_\text{total}}$ can be derived if the value of $P_{\max}$ is determined. Therefore, the global optimal solution of $P_{\max}$ is a key problem in this binary power control scheme.

When a candidate wireless subchannel CH_k is assigned into the maximum power transmission subchannel subset, the maximum transmission power $P_{\max\_1}$ used for the maximum power transmission subchannel subset is given by

$$P_{\max\_1} = \frac{P_{\max\_\text{total}}}{M}. \tag{8}$$

Furthermore, the energy efficiency of MIMO-OFDM communication system with the wireless subchannel CH_k is derived as

$$\eta_1 = \frac{\sum_{i=1}^{M} \log_2\left(1 + (P_{\max\_1}/n_0)\|H_i\|_F^2\right)}{P_{\max\_\text{total}}}. \tag{9}$$

When a candidate wireless subchannel CH_k is not assigned into the maximum power transmission subchannel subset, the maximum transmission power $P_{\max\_2}$ used for the maximum power transmission subchannel subset is given by

$$P_{\max\_2} = \frac{P_{\max\_\text{total}}}{M - 1}. \tag{10}$$

Furthermore, the energy efficiency of MIMO-OFDM communication system without the wireless subchannel CH_k is derived as

$$\eta_2 = \frac{\sum_{i=1, i\neq k}^{M-1} \log_2\left(1 + (P_{\max\_2}/n_0)\|H_i\|_F^2\right)}{P_{\max\_\text{total}}}. \tag{11}$$

Based on Principle 1, the candidate wireless subchannel CH_k can be finally assigned into the maximum power transmission subchannel subset and the maximum transmission power $P_{\max}$ can be derived under the condition of $\eta_1 \geq \eta_2$. From Principle 1, the optimization relationship of

maximum transmission power is derived in the Appendix and is expressed by

$$1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 \geq \frac{\prod_{i=1,i\neq k}^{M-1}\left(n_0+(P_{\max\_\text{total}}/(M-1))\|H_i\|_F^2\right)}{\prod_{i=1,i\neq k}^{M-1}\left(n_0+(P_{\max\_\text{total}}/M)\|H_i\|_F^2\right)}. \quad (12)$$

Compared with the transmission power over wireless subchannels, the value of AWGN n_0 is obviously less than the value of $(P_{\max\_\text{total}})/(M-1)\|H_i\|_F^2$ or $(P_{\max\_\text{total}}/M)\|H_i\|_F^2$. Therefore, the right side of (12) can be approximated as

$$\begin{aligned}1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 &\geq \frac{\prod_{i=1,i\neq k}^{M-1}\left(n_0+(P_{\max\_\text{total}}/(M-1))\|H_i\|_F^2\right)}{\prod_{i=1,i\neq k}^{M-1}\left(n_0+(P_{\max\_\text{total}}/M)\|H_i\|_F^2\right)} \\ &\approx \frac{\prod_{i=1,i\neq k}^{M-1}\left((P_{\max\_\text{total}})/(M-1)\|H_i\|_F^2\right)}{\prod_{i=1,i\neq k}^{M-1}\left((P_{\max\_\text{total}}/M)\|H_\text{i}\|_F^2\right)}.\end{aligned} \quad (13)$$

Furthermore, (13) can be derived as follows:

$$1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 \geq \frac{\prod_{i=1,i\neq k}^{M-1}(1/(M-1))}{\prod_{i=1,i\neq k}^{M-1}(1/M)}, \quad (14a)$$

$$\Downarrow$$

$$1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 \geq \frac{(1/(M-1))^{M-1}}{(1/M)^{M-1}}, \quad (14b)$$

$$\Downarrow$$

$$1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 \geq (M/(M-1))^{M-1}. \quad (14c)$$

Based on (14c), we can derive a current global optimization solution of maximum transmission power $P_{\max\_2}$ with the CSI of wireless communication system and the subchannel number of the maximum power transmission subchannel subset.

When the number of subchannels M approaches to infinite, we have the following result:

$$\lim_{M\to\infty}\left(\frac{M}{M-1}\right)^{M-1}=e. \quad (15)$$

From the numerical simulation of threshold values $(M/(M-1))^{M-1}$ and limitation values e in Figure 2, the difference between the threshold values $(M/(M-1))^{M-1}$ and limitation values e is less than 1.67% when the number of subchannel is large than or equal to 30. Therefore, we further derive two simple results for possible engineering applications.

Result 1. When the number of current subchannel in the maximum power transmission subchannel subset is less than 30, the value of maximum transmission power $P_{\max}$ is derived by

$$\frac{P_{\max}}{n_0}\|H_i\|_F^2 \geq \left(\frac{M}{M-1}\right)^{M-1}-1. \quad (16)$$

Result 2. When the number of current subchannel in the maximum power transmission subchannel subset is larger than or equal to 30, the value of maximum transmission power $P_{\max}$ are derived by

$$\frac{P_{\max}}{n_0}\|H_i\|_F^2 \geq e-1. \quad (17)$$

From above results, especially from Result 2, the complexity of maximum transmission power derivation can be reduced.

3.3. Algorithm Design. Based on Results 1 and 2, an energy efficiency binary power control (EEBPC) algorithm is designed for improving energy efficiency of MIMO-OFDM wireless communication systems. The detailed EEBPC algorithm is illustrated in Algorithm 1. Morever, an assumption of Algorithm 1 is that all subchannels of wireless subchannel set K are degressively ordered.

4. Simulation Results and Performance Analysis

Based on the new EEBPC algorithm, the energy efficiency and spectrum efficiency performance of MIMO-OFDM wireless communication systems is simulated and analyzed. In the following simulation, some parameters of the system model in Figure 1 are configured as follows: the total transmission power of base station is ranged from 0.6 to 1.4 watt (W); considering the OFDM scheme used in MIMO wireless communication system, the number of subchannels is ranged from 8 to 128; the AWGN n_0 in wireless subchannels is assumed as 0.1 W. Wireless subchannels are simulated by a Monte Carlo approach based on AWGN wireless subchannels with zero mean values.

From Figure 3, the impact of the number of subchannels on the spectrum efficiency of MIMO-OFDM communication system with the EEBPC algorithm is investigated. The spectrum efficiency of MIMO-OFDM communication system increases with the number of subchannels and the total transmission power of base station.

From Figure 4, the impact of number of subchannels on the energy efficiency of MIMO-OFDM communication system with the EEBPC algorithm is analyzed. The energy efficiency of MIMO-OFDM communication system increases with the number of subchannels, but the energy efficiency of MIMO-OFDM communication system decreases with the total transmission power of base station.

From Figure 5, the EEBPC algorithm is compared with the traditional average power control algorithm [16] in the spectrum efficiency of MIMO-OFDM communication system with different number of subchannels. Assume that the total transmission power of base station is configured as 1 W. When the number of subchannels is less than 12,

Input: $P_{\text{total}}, P_{\max\_\text{total}}$
Output: $M, P_{\max}, P_{\min}, K^{M}_{p\max}, K^{N-M}_{p\min}$
Initialization: Create a wireless sub-channel set K with N subchannels, the maximum power transmission subchannel subset $K^{M}_{p\max}$ and the minimum power transmission subchannel subset $K^{N-M}_{p\min}$,

$$K = \{\text{CH} \mid \text{CH} = \bigcup_{i=1}^{N} \text{CH}_i\},$$
$$K^{M}_{p\max} = \phi,$$
$$K^{N-M}_{p\min} = \phi,$$

Begin:
(1) Create a new set $\widetilde{K}$ from the set K by a descending order of $\|H_i\|_F^2$,

$$\widetilde{K} = \{\text{CH} \mid \text{CH} = \bigcup_{i=1}^{N} \text{CH}_i; \forall (1 \le i \le k \le N), \|H_i\|_F^2 \ge \|H_K\|_F^2\}$$

(2) **for** $i = 1 : N$ **do**

$$P_{\max} = \frac{P_{\max\_\text{total}}}{i-1}$$

if $i < 30$,
compare SNR_i with the threshold value,

$$\left(\text{SNR}_i = \frac{P_{\max}}{n_0}\|H_i\|_F^2\right) \ge \left(\frac{i}{i-1}\right)^{i-1} - 1$$

else
compare SNR_i with the imitation value,

$$\left(\text{SNR}_i = \frac{P_{\max}}{n_0}\|H_i\|_F^2\right) \ge e - 1$$

end if
if SNR_i large than or equal to the threshold or limitation values
add CH_i into $K^{M}_{p\max}$,
else
$M = i - 1$
add $\text{CH}_j (M + 1 \le j \le N)$ into $K^{N-M}_{p\min}$,
break,
end if
end for
(3)

$$P_{\max} = \frac{P_{\max\_\text{total}}}{M}, P_{\min} = \frac{P_{\text{total}-P_{\max\_\text{total}}}}{N-M},$$

end Begin

Algorithm 1: Energy efficiency binary power control.

the spectrum efficiency of MIMO-OFDM communication system with average power control algorithm is larger than that with EEBPC algorithm. When the number of subchannels is larger than or equal to 12, the spectrum efficiency of MIMO-OFDM communication system with EEBPC algorithm is larger than that with average power control algorithm. Moreover, the spectrum efficiency gain of EEBPC algorithm increases with the number of subchannels.

From Figure 6, the EEBPC algorithm is compared with the traditional average power control algorithm in the energy efficiency of MIMO-OFDM communication system with different number of subchannels. Assume that the total transmission power of base station is configured as 1 W. When the number of subchannels is less than 12, the energy efficiency of MIMO-OFDM communication system with average power control algorithm is larger than that with EEBPC algorithm. When the number of subchannels is larger than or equal to 12, the energy efficiency of MIMO-OFDM communication system with EEBPC algorithm is large than that with average power control algorithm. Moreover, the energy efficiency gain of EEBPC algorithm increases with the number of subchannels.

From Figure 7, the EEBPC algorithm is compared with the traditional average power control algorithm in the spectrum efficiency of MIMO-OFDM communication system with different total transmission power of base station. Assume that the number of subchannels is configured as 64. The spectrum efficiency of MIMO-OFDM communication system with EEBPC algorithm is large than that with average power control algorithm. Moreover, the spectrum efficiency gain of EEBPC algorithm is invariable with the total transmission power of base station.

From Figure 8, the EEBPC algorithm is compared with the traditional average power control algorithm in the energy efficiency of MIMO-OFDM communication system with different total transmission power of base station. Assume that the number of subchannels is configured as 64. The energy efficiency of MIMO-OFDM communication system with EEBPC algorithm is large than that with average power control algorithm. Moreover, the energy efficiency gain of

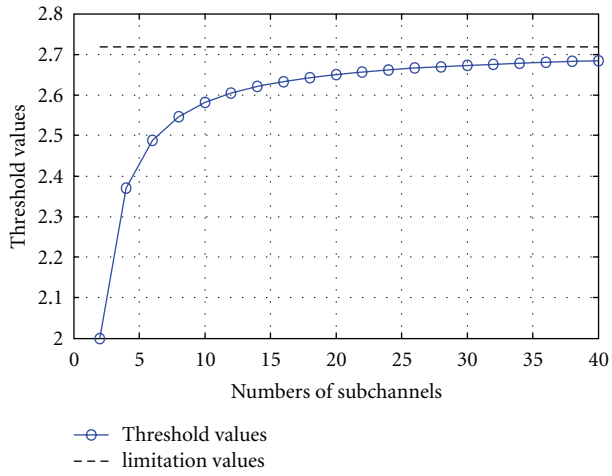


Figure 2: Threshold values versus number of subchannels.

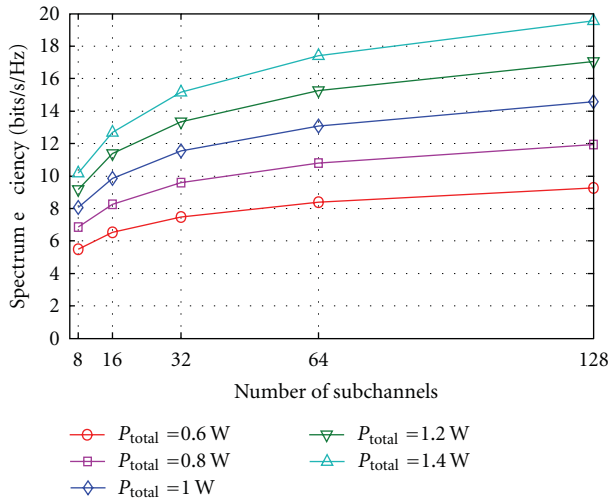


Figure 3: Spectrum efficiency analysis with number of subchannels limited by different total transmission power.

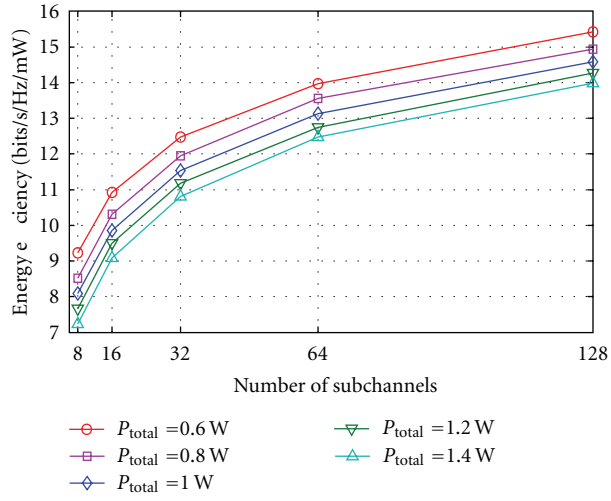


Figure 4: Energy efficiency analysis with number of subchannels limited by different total transmission power.

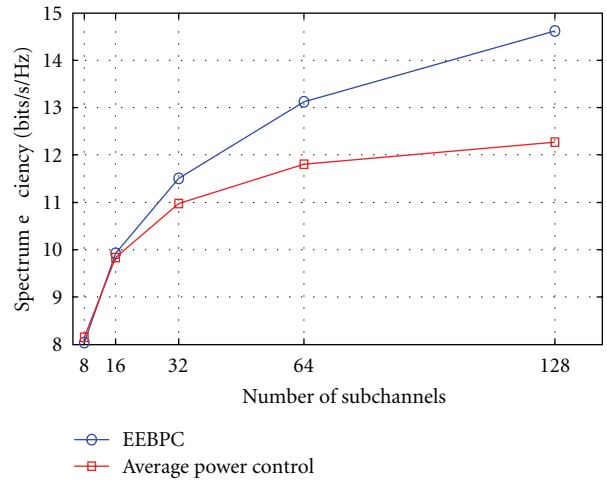


Figure 5: Comparison spectrum efficiency of EEBPC and average power control algorithms with different number of subchannels.

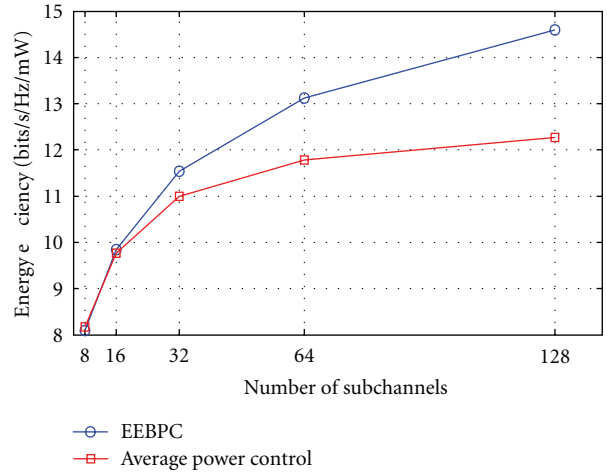


Figure 6: Comparison energy efficiency of EEBPC and average power control algorithms with different number of subchannels.

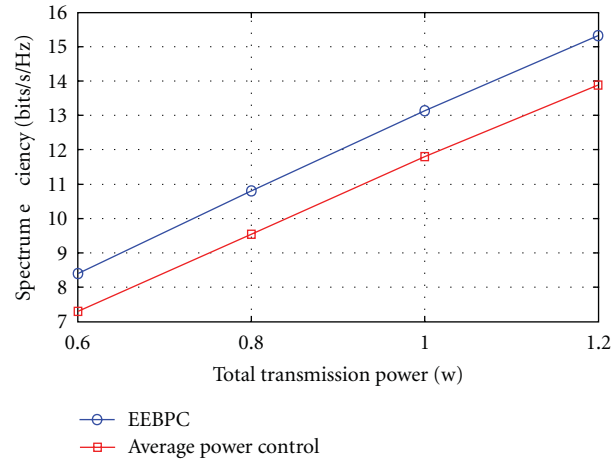


Figure 7: Comparison spectrum efficiency of EEBPC and average power control algorithms with different total transmission power.

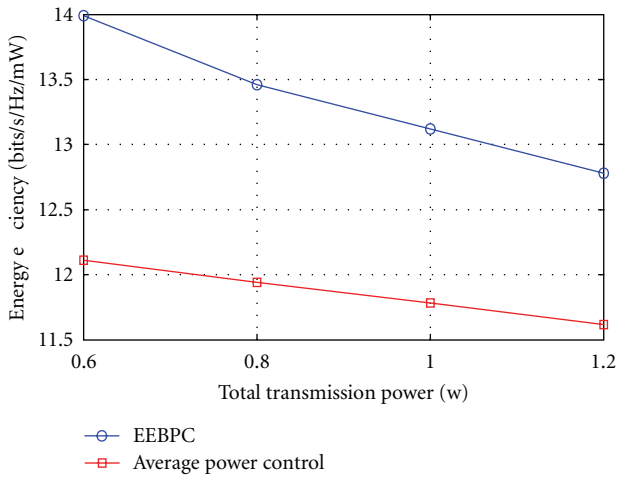


FIGURE 8: Comparison of energy efficiency of EEBPC and average power control algorithms with different total transmission power.

EEBPC algorithm decreases with the total transmission power of base station.

5. Conclusion

In this paper, the energy efficiency of binary power control scheme is investigated and formulated by three principles. Furthermore, a global optimalsolution of the maximum transmission power of MIMO-OFDM wireless communication system is derived. Moreover, a simple engineering application result is proposed for reducing the complexity of calculation. Based on them, a new EEBPC algorithm is designed for performance analysis. The exact impact of EEBPC algorithm on the spectrum efficiency and the energy efficiency has been fully investigated under different number of subchannels and total transmission power of base station. Simulation results have shown that the energy efficiency and spectrum efficiency of EEBPC algorithm is better than that of traditional average power control algorithm when the number of subchannels is larger than 11. Our future work includes a further investigation of the impact of multicell on the EEBPC algorithm.

Appendix

Appendix of (12)

In this appendix, we derive the optimization relationship of maximum transmission power. Based on Principle 1, only satisfying the condition of $\eta_1 \geq \eta_2$, the candidate wireless subchannel CH_k can be finally assigned into the maximum power transmission subchannel subset. Therefore, this condition is expressed by

$$\frac{\sum_{i=1}^{M}\log_2\left(1+(P_{\max\_1}/n_0)\|H_i\|_F^2\right)}{P_{\max\_\text{total}}} \geq \frac{\sum_{i=1,i\neq k}^{M-1}\log_2\left(1+(P_{\max\_2}/n_0)\|H_i\|_F^2\right)}{P_{\max\_\text{total}}}. \tag{A.1}$$

Based on (A.1), we can further derive this expression as follows:

$$\sum_{i=1}^{M}\log_2\left(1+\frac{P_{\max\_1}}{n_0}\|H_i\|_F^2\right) \geq \sum_{i=1,i\neq k}^{M-1}\log_2\left(1+\frac{P_{\max\_2}}{n_0}\|H_i\|_F^2\right),$$

$$\Downarrow \tag{A.2}$$

$$\log_2\prod_{i=1}^{M}\left(1+\frac{P_{\max\_1}}{n_0}\|H_i\|_F^2\right) \geq \log_2\prod_{i=1,i\neq k}^{M-1}\left(1+\frac{P_{\max\_2}}{n_0}\|H_i\|_F^2\right),$$

$$\Downarrow \tag{A.3}$$

$$\prod_{i=1}^{M}\left(1+\frac{P_{\max\_1}}{n_0}\|H_i\|_F^2\right) \geq \prod_{i=1,i\neq k}^{M-1}\left(1+\frac{P_{\max\_2}}{n_0}\|H_i\|_F^2\right). \tag{A.4}$$

Assume that the total transmission power of the maximum power transmission subchannel subset $P_{\max\_\text{total}}$ is fixed, so we can derive the following expression considering (8) and (10):

$$P_{\max\_2} \geq P_{\max\_1}. \tag{A.5}$$

Based on (A.5), and (A.4) is further derived as follows:

$$\left(1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2\right)\prod_{i=1,i\neq k}^{M-1}\left(1+\frac{P_{\max\_1}}{n_0}\|H_i\|_F^2\right) \geq \prod_{i=1}^{M}\left(1+\frac{P_{\max\_1}}{n_0}\|H_i\|_F^2\right) \geq \prod_{i=1,i\neq k}^{M-1}\left(1+\frac{P_{\max\_2}}{n_0}\|H_i\|_F^2\right),$$

$$\Downarrow$$

$$1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 \geq \frac{\prod_{i=1,i\neq k}^{M-1}\left(1+(P_{\max\_1}/n_0)\|H_i\|_F^2\right)}{\prod_{i=1,i\neq k}^{M-1}\left(1+(P_{\max\_2}/n_0)\|H_i\|_F^2\right)},$$

$$\Downarrow$$

$$1+\frac{P_{\max\_2}}{n_0}\|H_k\|_F^2 \geq \frac{\prod_{i=1,i\neq k}^{M-1}\left(n_0+(P_{\max\_\text{total}}/(M-1))\|H_i\|_F^2\right)}{\prod_{i=1,i\neq k}^{M-1}\left(n_0+(P_{\max\_\text{total}}/M)\|H_i\|_F^2\right)}. \tag{A.6}$$

This completes the derivation.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (NSFC), Contract/Grant number: 60872007, 61103177; National 863 High Technology Program of China, Contract/Grant number 2009AA01Z239; The Ministry of Science and Technology (MOST), China,

International Science and Technology Collaboration Program, Contract/Grant number 0903; EU FP7-PEOPLE-IRSES, project acronym S2EuNet, Contract/Grant number 247083.

References

[1] I Humar, X. Ge, X. Lin, M. Jo, and M. Chen, "Rethinking energy-efficiency models of cellular networks with embodied energy," *IEEE Network Magazine*, vol. 25, no. 2, pp. 40–49, 2011.

[2] W. Feng, Y. Li, S. Zhou, and J. Wang, "On power consumption of multi-user distributed wireless communication systems," in *Proceedings of the 2010 International Conference on Communications and Mobile Computing, (CMC 2010)*, pp. 366–369, China, April 2010.

[3] T. Kelly, "ICTs and climate change," Tech. Rep., ITU-T Technology, 2007.

[4] C.-X. Wang, X. Hong, X. Ge, X. Cheng, G. Zhang, and J. Thompson, "Cooperative MIMO channel models: a survey," *IEEE Communications Magazine*, vol. 48, no. 2, Article ID 5402668, pp. 80–87, 2010.

[5] H. Yang, "A road to future broadband wireless access: MIMO-OFDM-based air interface," *IEEE Communications Magazine*, vol. 43, no. 2, pp. 53–60, 2005.

[6] L. M. Correia, D. Zeller, O. Blume et al., "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Communications Magazine*, vol. 48, no. 11, Article ID 5621969, pp. 66–72, 2010.

[7] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 1, pp. 57–62, 1992.

[8] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 641–646, 1993.

[9] Y. H. Lin and R. L. Cruz, "Power control and scheduling for interfering links," in *Proceedings of the 2004 IEEE Information Theory Workshop*, pp. 288–291, San Antonio, Tex, USA, October 2004.

[10] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 145–152, 2004.

[11] S. de la Kethulle de Ryhove, G. E. Oien, and F. Bohagen, "Maximizing the average spectral efficiency of dual-branch MIMO systems with discrete-rate adaptation," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 1003–1011, 2009.

[12] H. S. Kim and B. Daneshrad, "Energy-constrained link adaptation for MIMO OFDM wireless communication systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 9, Article ID 5510775, pp. 2820–2832, 2010.

[13] B. Zayen, M. Haddad, A. Hayar, and G. E. Oien, "Binary power allocation for cognitive radio networks with centralized and distributed user selection strategies," *Physical Communication Journal*, vol. 1, no. 3, pp. 183–193, 2008.

[14] A. Gjendemsjo, D. Gesbert, G. E. Oien, and S. G. Kiani, "Binary power control for sum rate maximization over multiple interfering links," *IEEE Transactions on Wireless Communications*, vol. 7, no. 8, Article ID 4600228, pp. 3164–3173, 2008.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 545–554, 623—656, 1948.

[16] Z. Zhihua, X. He, and W. Jianhua, "Average power control algorithm with dynamic channel assignment for TDD-CDMA systems," in *Proceedings of the 2008 International Conference on Advanced Infocomm Technology*, Shenzhen, China, July 2008.

[17] J. W. Cho, J. Mo, and S. Chong, "Joint network-wide opportunistic scheduling and power control in multi-cell networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, Article ID 4801504, pp. 1520–1531, 2009.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 212737, 11 pages
doi:10.1155/2011/212737

Review Article

A Survey of Adaptive and Real-Time Protocols Based on IEEE 802.15.4

Feng Xia, Ruonan Hao, Yang Cao, and Lei Xue

School of Software, Dalian University of Technology, Dalian 116620, China

Correspondence should be addressed to Feng Xia, f.xia@ieee.org

Received 2 April 2011; Accepted 4 September 2011

The IEEE 802.15.4 standard is a protocol that has been widely used in many applications exploiting wireless sensor networks (WSNs). However, it does not provide any means of differentiated services to improve the quality of service (QoS) for time-critical and delay-sensitive events. Furthermore, adaptive throughput performance for individual nodes cannot be supported with the current specifications. In this survey paper, we first discuss the negative aspects of IEEE 802.15.4 MAC in contention access period, contention-free period, and the overall cross-period, respectively, in terms of adaptive and real-time guarantees. We then give an overview on some interesting mechanisms used in existing adaptive and real-time protocols in compliance with IEEE 802.15.4. Careful examination of such research works reveals that by optimizing the original specifications and dynamically adjusting the protocol parameters, the total network efficiency can be significantly improved. Nevertheless, there are still certain challenges to overcome in pursuing the most appropriate protocol without introducing any unacceptable side effects.

1. Introduction

Wireless sensor networks (WSNs) are being deployed in a variety of real-life applications, such as factory automation [1], distributed and process control [2, 3], real-time monitoring of machinery health [4, 5], detection of liquid/gas leakage, and radiation check [6]. This has been made possible by the adoption of two industrial standards: (1) IEEE 802.15.4 standard [7], which defines the physical and medium access control (MAC) layers of the protocol stack; and (2) the ZigBee specification [8], which covers the networking and application layers.

IEEE 802.15.4, in particular, is the emerging next-generation standard uniquely designed for low-rate wireless personal area networks (LR-WPANs) [7]. And it has become one of the most popular communication protocols for wireless interconnection of fixed/or portable devices [9]. Nevertheless, the current specifications of IEEE 802.15.4 lack quality of service (QoS) mechanisms that are required for adaptive and real-time WSN applications. Actually, it has been known that WSNs based on IEEE 802.15.4 suffer from severe adaptivity and timeliness issues, especially when power management is enabled for conserving energy.

In many WSN applications, different data packets may have different importance depending on the information they contain. For example, an alarm message should have priority over a packet with noncritical sensor readings. Unfortunately, IEEE 802.15.4 treats all packets in the same way, which may result in unfairness and degradation of the network performance, particularly in high-load conditions. In addition, the MAC design is inadequate to deal with when certain nodes are sending data more frequently compared to others. Also, the traffic and network conditions in a WSN are usually very dynamic because of the noisy wireless channel and the failure probability of sensor nodes. Hence, data transmissions should be flexible enough to support a wide variety of scenarios to adapt to the actual operating conditions. All of the above requirements make the original MAC design of IEEE 802.15.4 inadequate to achieve network efficiency. Providing QoS support in WSNs for improving their timeliness and adaptivity performance under severe energy constraints has attracted tremendous research works [10–12] recently. Lots of adaptive and real-time protocols based on IEEE 802.15.4 have been proposed. In an adaptive and real-time protocol, each node is expected to perform real-time computations and to send high-quality data with guaranteed QoS [13]. Communications must be both adaptive and conducted on time. In fact, the standard IEEE 802.15.4 MAC protocol shows great potential for flexibly fitting different requirements of WSN applications by adequately setting its

parameters (low-duty cycles, guaranteed time slots (GTS)) [14].

In this survey paper, we will talk about limitations of original IEEE 802.15.4 MAC protocol in terms of contention access period (CAP), contention-free period (CFP), and the whole in detail. Further, a number of adaptive and real-time protocols based on IEEE 802.15.4 are discussed. According to these existing research works, we finally present an in-depth analysis of current challenges and technological trends, and possible solutions are predicted.

The remainder of this paper is organized as follows. Section 2 gives an overview of IEEE 802.15.4 standard specifications. In Section 3, we will discuss major features of IEEE 802.15.4 MAC in CAP that limit network efficiency, and look at current approaches to overcome these constraints. Further analysis and discussions of such approaches are also presented in this section. Sections 4 and 5 talk about similar issues in CFP and overall cross-period, respectively. Finally, Section 6 concludes this paper.

2. Overview of IEEE 802.15.4 Medium Access Control

According to the standard IEEE 802.15.4 [7], it defines two different channel access methods: beacon-enabled mode and nonbeacon-enabled mode.

In beacon-enabled mode, a PAN is formed by a PAN coordinator, which is in charge of managing the whole network, and optionally, by one or more coordinators which are responsible for a subset of nodes in the network. beacon frames are periodically sent by the PAN coordinator to identify its PAN and synchronize nodes that are associated with it. The PAN coordinator defines a superframe structure characterized by a *beacon interval* (BI) and a *superframe duration* (SD). BI specifies the time between two consecutive beacons, and consists of an active period and, optionally, an inactive period. The active period, also called superframe, is corresponding to SD and can be divided into 16 equally sized time slots, during which frame transmissions are allowed. While during the inactive period, all nodes may enter a low-power state to save energy. The superframe structure of beacon-enabled mode is depicted in Figure 1.

BI and SD are determined by two parameters, the *beacon order* (BO) and the *superframe order* (SO), respectively, which are broadcasted by the coordinator via a beacon to all nodes. They are defined as follows:

$$\begin{aligned} \text{BI} &= \text{aBaseSuperframeDuration} * 2^{\text{BO}} \\ \text{SD} &= \text{aBaseSuperframeDuration} * 2^{\text{SO}} \\ &\text{for } 0 \le \text{BO} \le \text{SO} \le 14. \end{aligned} \tag{1}$$

SD can be further divided into a CAP and a CFP. During the CAP, a slotted CSMA/CA (carrier sense multiple access with collision avoidance) algorithm is used for channel access. While in the CFP, communication occurs in a TDMA (time division multiple access) style by using a number of GTSs, preassigned to individual nodes. It is also worth pointing out that the nonbeacon-enabled mode is entirely contention based. There is no superframe and nodes are always active. In this paper, we only select the more commonly used beacon-enabled mode as a basic operating mode in consideration.

2.1. CSMA/CA. The slotted CSMA/CA backoff algorithm mainly depends on three variables: (1) the *backoff exponent* (BE) enables the computation of the backoff delay, which is the slot periods a device must wait before attempting to access the channel; (2) the *contention window* (CW) represents the number of backoff periods during which the channel must be sensed idle before the transmission can start; (3) the *number of backoffs* (NB) represents the number of times the CSMA/CA algorithm is required to backoff while attempting to access the channel.

Upon receiving a data frame to be transmitted, the slotted CSMA/CA algorithm performs the following steps.

Step 1. A set of state variables is initialized: NB = 0, CW = 2, and BE = *macMinBE*. If the battery life extension variable is set, the maximum value of BE can be only 2.

Step 2. The MAC layer delays for a random number of complete slot periods in the range 0 to $2^{\text{BE}} - 1$, which is generated to initialize a backoff timer.

Step 3. After the completion of the backoff periods, a clear channel assessment (CCA) is performed to check the state of the wireless medium.

Step 4. If the channel is busy, the state variable are updated as follows: NB = NB + 1, BE = Min(BE + 1, *macMaxBE*), and CW = 2. If the maximum number of backoffs (NB = *macMaxCSMABackoffs* = 5) is reached, the algorithm shall terminate with a channel access failure status. Otherwise, it falls back to *Step 2*.

Step 5. If the channel is assessed to be idle, the MAC sublayer must ensure that the contention window is expired before starting transmission. For this, the MAC sublayer first decrement CW by one. If CW = 0, then the frame is transmitted. Otherwise, the algorithm must go back to *Step 3* to perform a second CCA. Figure 2 represents the above steps graphically.

2.2. GTS Mechanism. Upon receiving the request, the coordinator first checks the availability of GTS slots in the current superframe, based on the remaining length of the CAP and the desired length of the requested GTS. Provided there is sufficient capacity in the current superframe, the coordinator determines, based on a first come first serviced (FCFS) fashion, a device list for GTS allocation in the next superframe, and informs the device about the allocation of slot in the GTS descriptor in the following beacon frame.

GTS deallocation can be performed by the coordinator or by the device itself. For device initialized deallocation, it sends GTS request with characteristic-type subfield set to zero using CSMA/CA during CAP. However, in most cases, the PAN coordinator has to detect the activities of the devices

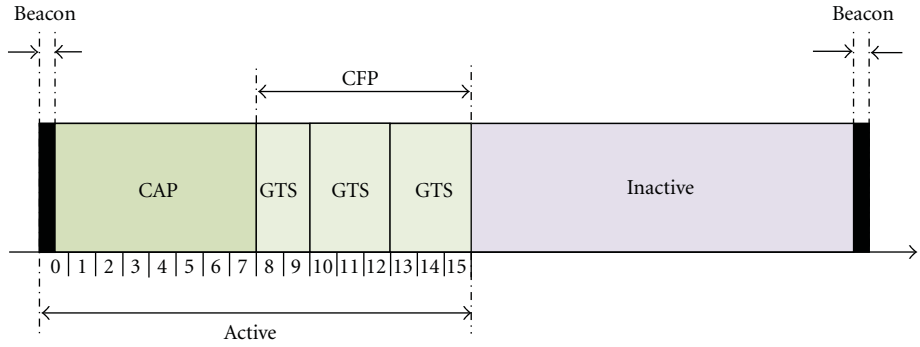


Figure 1: IEEE 802.15.4 superframe structure.

occupying GTSs and determine when the devices stop using their GTSs. If the coordinator does not receive data from the device in the GTS for at least 2*n superframes, the coordinator will deallocate the GTS with starting slot subfield set to zero in the GTS descriptor field of the beacon frame for that device. The value of n is defined as follows:

$$n = 2^{8-\mathrm{BO}}, \quad 0 \le \mathrm{BO} \le 8,$$
$$n = 1, \quad 9 \le \mathrm{BO} \le 14. \tag{2}$$

3. Approaches for Contention Access Period

In beacon-enabled mode, IEEE 802.15.4 medium access control is ruled by the slotted CSMA/CA mechanism in the contention access period. However, the standard slotted CSMA/CA mechanism does not provide any means of differentiated services to improve the quality of service for time-critical events (such as alarms, PAN management messages, and time slot reservation). Also, the IEEE 802.15.4 MAC layer cannot support different throughput performance for individual nodes with the current specifications. If certain nodes are sending data more frequently compared to others, with the standard slotted CSMA/CA mechanism, it is hard to achieve network efficiency. Numerous results have shown that the CSMA/CA-based 802.15.4 MAC has a very poor performance in terms of adaptivity and real-time guarantees, especially when a large number of sensor nodes start transmitting simultaneously [15–17]. Thus, it is very important to understand the fundamental reasons of this limitation in order to mitigate its negative impact.

The behavior of slotted CSMA/CA is affected by four initialization parameters, which are (1) the minimum backoff exponent (*macMinBE*); (2) the maximum backoff exponent (*macMaxBE*); (3) the initial value of CW ($\mathrm{CW}_{\mathrm{init}}$); (4) the maximum number of backoffs (*macMaxCSMABackoffs*).

For *macMinBE* and *macMaxBE*: the IEEE 802.15.4 standard provides a fixed backoff range which limits the default value of *macMinBE* as 3, and *macMaxBE* as 5, giving only the range of $[2^3, 2^5]$ for randomly selecting the actual backoff value. Changing the value of *macMinBE* or *macMaxBE* will have an impact on the network performance. For example, if the BE value is decremented to a value less than the default value 3, the lower boundary of the possible backoff value will decrease also. It will shorten the waiting time when CCA detects the channel busy or when a packet collides with a different packet in the channel. This gives a higher chance of selecting a shorter backoff time, and makes the node try the CCA more frequently, leading it to a higher possibility of making a successful transmission. As a result, the throughput will increase significantly compared to other nodes with a longer waiting time.

A performance evaluation study in [9] also shows that the network throughput is independent from the initial value of the backoff exponent *macMinBE* for a large-scale WSN, which is because the lower limit of the backoff delay interval $[0, 2^{\mathrm{BE}} - 1]$ is not affected by the choice of *macMinBE*. However, the impact of *macMinBE* on the network throughput is quite important in small-scale networks. In fact, increasing *macMinBE* will lead to relatively lower network throughput (since the capacity of the network is not entirely used for high *macMinBE*), but to significant higher success probability thanks to more efficient collision avoidance.

For $\mathrm{CW}_{\mathrm{init}}$, the initial value of CW is another medium access parameter that is useful to differentiate packet transmissions in slotted CSMA/CA of IEEE 802.15.4. It represents the number of CCAs, which is performed prior to each packet transmission to determine whether the medium is busy or idle.

The standard specifies that the transmitter node performs the CCA twice in order for the purpose of protecting acknowledgement (ACK) frame and giving enough time for a receiving device to process the frame. When an ACK frame is acquired by the transmitter, the receiver should send it after t_{ACK} time which varies from 12 to 31 symbols (one backoff period is 20 symbols). Hence, one time of CCA can potentially cause a collision between a newly transmitted packet and an ACK packet. Nevertheless, one time of CCA gives a strong priority to obtain the channel. For a backoff counter, the range of the counter drawn for a high-priority packet's backoff procedure should be bounded by a smaller constant value than a normal packet's in order to enable a faster transmission. Provided that ACK collision is not so severe, the flexible CW value will definitely guarantee a better rate of successful transmission for a device with high priority compared with the conventional IEEE 802.15.4 MAC protocol, where every device initiates the CCA procedure with the same CW.

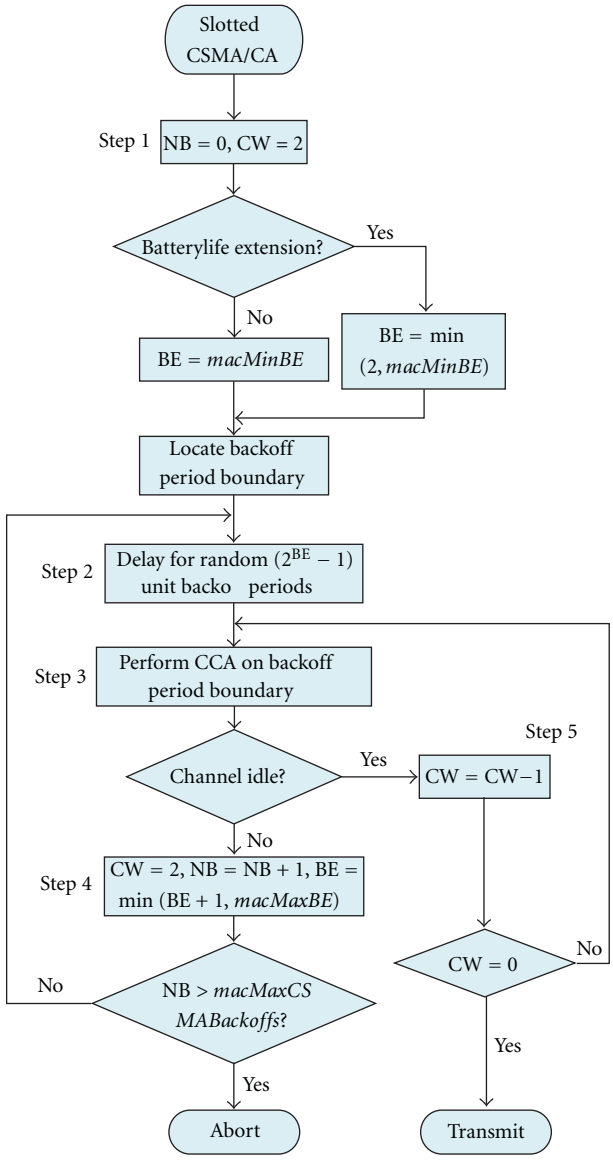


Figure 2: Slotted CSMA/CA algorithm.

The parameter *macMaxCSMABackoffs*, which represents the maximum number of times the CSMA/CA algorithm is required to backoff while attempting to access the channel, also serves to affect network performance for IEEE 802.15.4. In an ideal communication environment, nearly all undelivered packets are dropped by the protocol because they exceed the maximum number of backoff stages (i.e., *macMaxCSMABackoffs*). In this way, a larger *macMaxCSMABackoffs* which means a larger number of CSMA backoffs, can result in more packets that can be transmitted successfully and lead to a lower packet loss rate. As Figure 3 shows [18], increasing the *macMaxCSMABackoffs* parameter, while leaving all other parameters to their default values, results in an almost linear increase in the delivery ratio. However, the end-to-end delay will increase with the value of *macMaxCSMABackoffs*. This is because a larger *macMaxCSMABackoffs* value implies a larger number of packets is

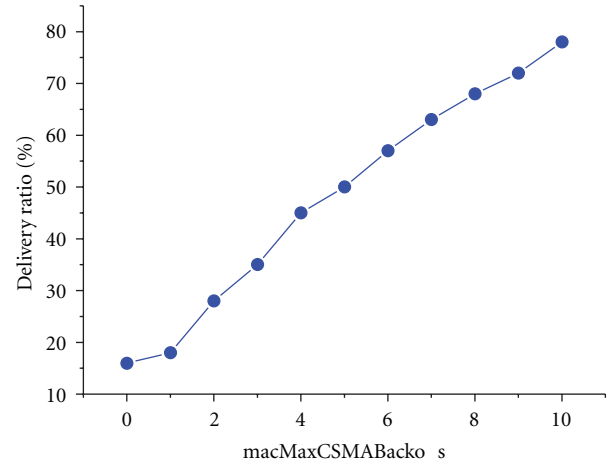


Figure 3: Impact of the *macMaxCSMABackoffs* on the delivery ratio.

successfully transmitted, which takes a longer backoff time, and in turn may cause longer end-to-end delay.

These impact of MAC parameters on the performance of 802.15.4 PAN suggests that the MAC low-efficiency problem, which is originated by the CSMA/CA algorithm, is made worse by the default parameter settings, which appears to be inappropriate for most WSN applications. Hence, the key question to answer is whether a more appropriate parameter setting can solve the problem without introducing unacceptable side effects. Recently, a wide variety of parameter tuning approaches in CAP have been proposed to improve network efficiency.

3.1. Adaptive Backoff Exponent Mechanism. Koubaa et al. provide a service differentiation mechanism in [13], particularly based on the *macMinBE* and *macMaxBE* parameters. Due to different importance of data traffic and command traffic, the particular mechanism considers command frames as the high-priority service class and data frames as the low-priority service class. The differentiated service strategies are presented in Figure 4. Instead of having the same CSMA/CA parameters for both traffic types, the algorithm assigns each class its own attributes, and denote [$macMinBE_{HP}$, $macMaxBE_{HP}$] and CW_{HP} the backoff interval and the contention window initial values for high-priority traffic related to command frames, and [$macMinBE_{LP}$, $macMaxBE_{LP}$] and CW_{LP} the initial values for low-priority traffic related to data frames.

By setting CW_{HP} higher than CW_{LP}, it results that low-priority traffic has to assess the channel to be idle for a longer time before transmission. And providing lower backoff delay values for high-priority traffic by setting $macMinBE_{HP}$ lower than $macMinBE_{LP}$ would improve its responsiveness without degrading its throughput.

In addition to the specification of different CSMA/CA parameters, priority queuing is applied to reduce queuing delays of high-priority traffic (Figure 4). In this case, slotted CSMA/CA uses priority scheduling to select frames from

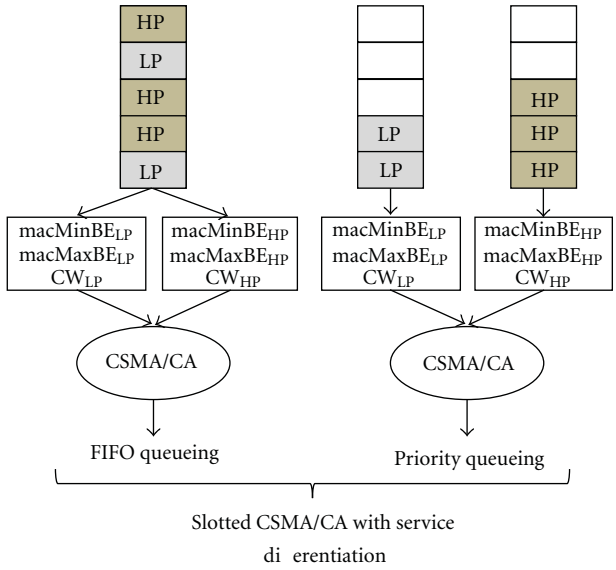


Figure 4: Differentiated service strategies.

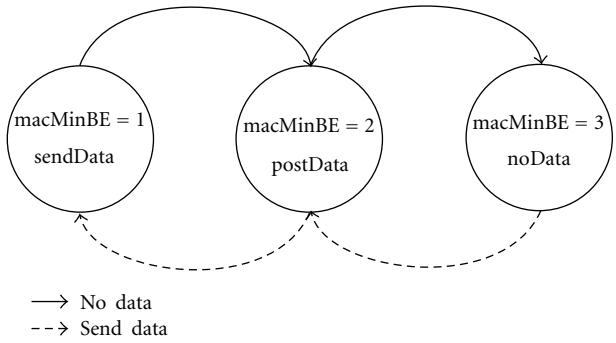


Figure 5: State transition scheme.

queues, and then applies the adequate parameters corresponding to each service class.

This differentiated service scheme for slotted CSMA/CA in IEEE 802.15.4 serves to improve the performance of time-sensitive messages. And it has been shown that tuning adequately the BE parameter of slotted CSMA/CA may result in an improved quality of service for time-critical messages.

A new state transition scheme is additionally proposed in [14]. By adjusting the *macMinBE* value of some nodes to a smaller value and by dynamically changing the value depending on the transmission conditions, the scheme shortened the backoff delay of nodes with frequent transmission.

In the MAC modifications, the value of the *macMinBE* is a number between 1 and 3, which changes flexibly as the condition of the node changes. As seen in Figure 5, each node has three states, *noData*, *postData*, and *sendData*, with each state having the default *macMinBE* value of 3, 2, and 1, respectively.

In the state transition scheme, The three states are switched dynamically and the nodes are requested to count the number of idle beacon frames (with no transmission) and the number of successful transmissions within a beacon frame, before the CCA process. By counting the numbers we can allow the nodes with more data to transmit to a higher priority in the network. The state transition scheme will not make additional fairness problems because if a node has nothing to send any more it increases the *macMinBE* so that it can be excluded from the high-priority nodes.

The change lets the modified node take advantage in transmitting data compared to the nonmodified nodes, causing higher throughput performance for the modified node. By implementing the state transition scheme, the throughput performance of the overall network is significantly improved.

Also, Rao and Marandin present a brief study of the CSMA/CA mechanism in [19], with emphasis on the improper BE distribution which results in frequent packet collisions and a loss in systems performance. And they additionally provided an algorithm called the adaptive backoff exponent (ABE) which reduces the probability of devices choosing identical number of backoff periods at collision rates, thus improving the system's performance considerably at these rates.

The ABE algorithm is primarily based on three important principles. Firstly is the idea of providing a higher range of backoff exponents to the devices, to reduce the probability of devices choosing the same number of backoff periods to sense the channel. Secondly is to do away with a constant minimum backoff exponent (*macMinBE*) value as used in the standard CSMA/CA. In this algorithm, the minimum backoff exponent shall be variable, hence devices are not likely to start off with the same backoff exponent when they wish to start a data transmission. And thirdly is the way the minimum backoff exponent is maintained. Since the algorithm implements a variable *macMinBE*, the variation factor is each node's contribution to the network traffic. Only devices that are involved in a transmission are taken into consideration. And devices that are not transmitting do not come under the purview of the algorithm.

According to this algorithm, all devices that are contributing more to the network traffic are slapped with higher *macMinBEs*, and devices which contribute less to the network congestion will use lower minimum backoff exponents. Therefore devices with higher *macMinBE* values are likely to wait longer than devices with lower *macMinBEs*, leading to an overall improvement in effective data bandwidth.

3.2. Adaptive Contention Window Mechanism. A frame tailoring (FRT) strategy is proposed in [20] to avoid ACK and data packet collision while allowing one-time CCA so that it can be exploited to provide strong prioritization in addition to the standard CSMA/CA.

In this scheme, the length of t_{ACK} is determined as depicted in Figure 6, depending on packet length, and the term frame tail is defined as the length of the remainder after the total packet length is divided by the backoff slot length (i.e., 20 symbols). If a frame tail is from 0 to 8 symbols, a receiver transmits an ACK packet at the very next backoff slot boundary as depicted in Figures 6(a) and 6(b). On the other hand, if a frame tail ranges from 9 to 19 symbols, an ACK transmission by the receiver is postponed to one backoff slot after the next backoff slot boundary to allow

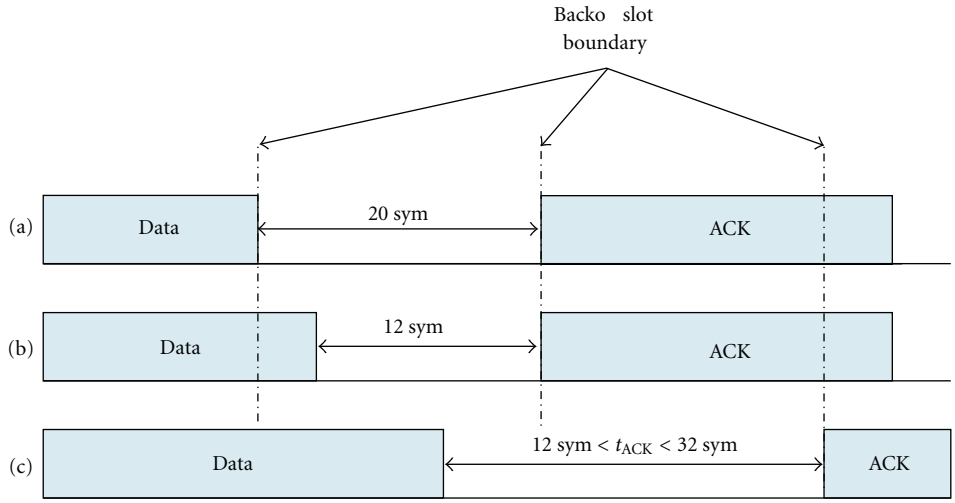


Figure 6: Variable t_{ACK} depending on the data packet length.

adequate time to prepare the ACK transmission. As a result, t_{ACK} becomes more than 20 symbols as shown in Figure 6(c). Then, the CCA operations of other contending nodes during this time interval report that the medium is idle. To protect ACK transmissions in such cases, two-time CCAs are mandated in IEEE 802.15.4 slotted CSMA/CA. FRT strategy is to adjust each data packet length so that t_{ACK} becomes exactly 12 symbols as shown in Figure 6(b). By doing so, one-time CCA will never declare an idle medium during the time period between a data and an ACK, and hence by adopting one-time CCA for a particular transmission, high prioritization can be achieved.

The proposed frame tailoring strategy effectively separates the medium access of each group of packet transmissions according to packet's priority. By adopting the proposed scheme, the probability of transmission deferment to the next active period due to competitive contention is relaxed and bounded delay is provided to high-priority packets.

Furthermore, Kim and Kang proposed a mechanism of contention window differentiation (CWD) in [21] to provide multilevel differentiated services for IEEE 802.15.4 sensor networks. CWD is a mechanism assigning various values of CW according to the priority classes. Let $class0, \ldots, classQ$ be the set of priority classes ordered by

$$classQ \prec class(Q-1) \prec \cdots \prec class0, \tag{3}$$

where $\prec$ denotes the order of priority. Equation (3) implies that $class0$ and $classQ$ are the highest and lowest priority classes, respectively. They differentiate the corresponding CW value of priority $classQ$ by $\mathrm{CW}[Q]$ as follows:

$$\mathrm{CW}[0] \leq \mathrm{CW}[1] \leq \cdots \leq \mathrm{CW}[Q]. \tag{4}$$

The relationship in (4) is intuitive, since a device with a smaller CW has a better chance of transmission than a device with a larger CW in general. In other words, a device with high priority can start transmission when a device with low priority is performing the CCA procedure. It guarantees a better rate of successful transmission for a device with high priority compared with the conventional 802.15.4 MAC protocol, where every device initiates the CCA procedure with the same CW.

Numerical results show that the IEEE 802.15.4 standard in WSNs can support adaptive and timely packet transmissions by tuning the MAC parameters to more appropriate values. However, the increase in reliability is usually achieved at the cost of a higher latency, and high adaptivity and low delay may demand a significant energy consumption and network complexity, thus making a great many approaches not feasible at best. And due to the random nature of CSMA/CA algorithm, an appropriate parameters setting which guarantees both adaptivity and bounded latency for real-time applications is hardly achieved. There are also other challenges, it is not clear how to adapt the parameters to the changes of network and traffic regimes by algorithms that can run on resource-constrained nodes. A simple and accurate model of the influence of these parameters on the success probability, real-time performance, and adaptivity to various conditions as a whole is not available. What is worse, the cost to be paid, in most cases, will turn out to be even higher in a real environment.

The above discussed approaches and analysis pave the way for further research. Since most appropriate MAC parameters should depend on real operating conditions and specific QoS requirements, the ideal adaptive and real-time approach for CAP should dynamically select appropriate parameters to offer the required QoS support according to various operating conditions.

4. Approaches for Contention-Free Period

Based on the standardized IEEE 802.15.4 protocol, timeliness guarantee and adaptive throughput are the most important features that we have to pay attention to. Besides, timeliness guarantee is also appealing to WSN applications. As the

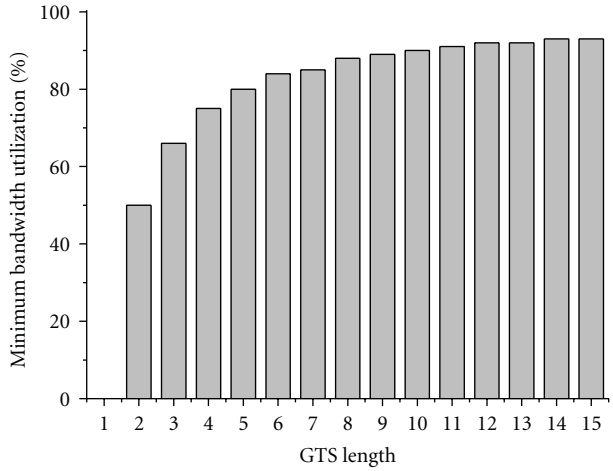


Figure 7: Minimum utilization limits of an explicit allocation.

requirements of WSNs, low data rate, low power consumption, and low cost wireless networking becomes more and more outstanding recently. Therefore the IEEE 802.15.4 protocol also provides real-time guarantees by using the GTS mechanism. This feature is quite attractive for time-sensitive WSNs. In fact, when operating in beacon-enabled mode, that is, beacon frames are transmitted periodically by a central node called PAN coordinator for synchronizing the network, the IEEE 802.15.4 protocol allows the allocation/deallocation of GTSs in a superframe for nodes that require real-time guarantees. Hence, the GTS mechanism provides a minimum service guarantee for the corresponding nodes and enables the prediction of the worst-case performance for each node's application.

However, the GTS mechanism also presents several negative impacts.

(1) It presents some limitations in terms of efficiency and deployment with a large number of nodes.

(2) because only up to seven GTSs (1 up to 15 time slots per GTS) can be allocated during each superframe, the GTSs can be quickly consumed by a few number of nodes, preventing the others from having a guaranteed service.

(3) A node with a low arrival rate that has been allocated a GTS may use it only partially (when the amount of guaranteed bandwidth is higher than its arrival rate). This leads to underutilization of the GTS bandwidth resources.

Now, for a CFP of a length k time slots, the minimum utilization limit is defined as follows in [22]:

$$\left(U_{\min}^{k}\right) = \frac{k-1}{k}, \quad 1 \le k \le 15. \tag{5}$$

Figure 7 presents the minimum utilization limits for different GTS length values, for one node. From Figure 7, it can be understood that the lowest utilizations can be experimented for GTSs with one time slot allocation. This is because the arrival rates of the flows can be low fractions of the indivisible R_{TS} (defined as the guaranteed bandwidth per one time slot), which triggers the motivation for sharing the time slot with other nodes, if the delay requirements of the flows can still be satisfied. This case is most likely to happen in sensor networks since their arrival rates may be particularly low.

In order to overcome the previously described limitations of the explicit GTS allocation in the IEEE 802.15.4 protocol, a great number of effective approaches have been developed. In this paper, we mainly focus on four popular adaptive and real-time approaches for CFP.

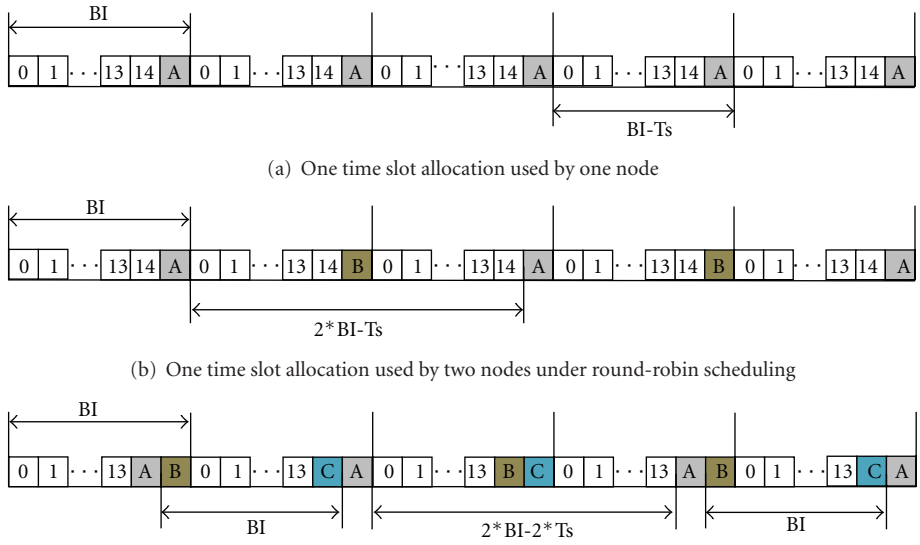
4.1. Adaptive GTS Allocation Scheme (AGA). Huang et al. proposed an energy-efficient protocol [23] that would be superior to the explicit GTS allocation mechanism. The AGA mechanism relies on assigning priorities in a dynamic fashion based on recent GTS usage feedbacks with the consideration of low latency and fairness. An ideal GTS allocation scheme has a good estimate of the future GTS usage behaviors of devices. With the estimate, the PAN coordinator allocates GTS resources to needy devices and reclaims the previously allocated but unused GTSs.

To achieve the above goal, the AGA mechanism arranges two phases in the scheme. In the classification phase, devices are assigned priorities in a dynamic fashion based on recent GTS usage feedbacks. Devices that need more attention from the coordinator are given higher priorities. In the GTS scheduling phase, GTSs are given to devices in a nondecreasing order of their priorities. A starvation avoidance mechanism is presented to regain service attention for lower-priority devices that need more GTSs for data transmissions.

The AGA scheme is developed based on the standard of the IEEE 802.15.4 MAC protocol and completely follows the specification defined in [7] without introducing any extra protocol overhead. Therefore the priority number of a device reflects its long-term transmission characteristics, that is, the scheme provides a multilevel AIMD [24] algorithm for updating the priority numbers. And the scheduling criteria are based on the priority numbers, the superframe length, and the GTS capacity of the superframe. And the numerical results indicate that the AGA scheme greatly outperforms the existing implementations.

4.2. Implicit Allocation Mechanism (i-Game). Based on the basic idea of sharing the same GTS by multiple flows, Koubaa et al. proposed the implicit allocation mechanism [22]. The allocation is based on implicit GTS allocation requests, taking into account the traffic specifications and the delay requirements of the flows.

While the GTS allocation mechanism is based on the traffic specification of the requesting nodes, their delay requirements, and the available GTS resources [25]. Instead of asking for affixed number of time slots, a node that wants to have a guaranteed service sends its traffic specification and delay requirement to the PAN coordinator. Then, the latter runs an admission control algorithm based on this information and the amount of available GTS resources.



(a) One time slot allocation used by one node

(b) One time slot allocation used by two nodes under round-robin scheduling

(c) Two time slot allocation used by three nodes under round-robin scheduling

Figure 8: Different implicit GTS allocations.

The new allocation request will be accepted if there is a schedule that satisfies its requirements and those of all other previously accepted allocation requests; otherwise, the new allocation request is rejected. The *i*-Game has the advantage of accepting multiple flows sharing the same GTS, while still meeting their delay requirements. It also improves the utilization of the CFP by reducing the amount of wasted bandwidth of GTSs and maximizes the duration of the CAP, since the CFP length is reduced to a minimum. With the help of network calculus, this GTS mechanism shows how to fairly share the allocation of k time slots in the CFP between N requesting nodes, with respect to their flow specifications. Observe in Figure 8 that changing the scheduling policy results in a change of the service curve, even if the guaranteed bandwidth is the same.

4.3. Knapsack Algorithm. A knap problem is formulated to obtain optimal GTS allocation such that a minimum bandwidth requirement is satisfied for the sensor devices. Hanson et al. in [26] have already shown that the knapsack scheme can achieve better GTS utilization and higher packet delivery ratio than the standard IEEE 802.15.4 scheme does.

The main object of the proposed knapsack algorithm is to improve the GTS allocation scheme in the IEEE 802.15.4-based MAC when used for a large number of medical and physical sensor devices deployed in a wireless body area sensor network (WiBaSe-Net) [27]. Shrestha et al. also proposed an optimization model, which takes the priority that is based on the packet generation rate of each device into account. In this allocation model, it assumes that if a device does not send GTS request or misses the beacon frame, it can use slotted CSMA/CA to transmit its data. If the request is unsuccessful, the device waits for the next beacon to send another GTS request. If the packet waiting time exceeds this delay limit, the sensor device simply discards the packet. The coordinator collects all the GTS requests during CAP and solves the knapsack algorithm for GTS allocation before transmitting the beacon frame. It saves the remaining bandwidth that is not allocated for GTS to use in the next super frame. That is the advantage that with the minimum bandwidth requirement the sensor devices can still meet their needs.

4.4. GTS Scheduling Algorithm (GSA). Na et al. proposed the GSA algorithm [28], which differs from the existing algorithms in that it is an on-line scheduling algorithm and allows transmissions of bursty and periodic messages with time constraints even when the network is overloaded. The evaluation of GSA mechanism is up to 100% higher than the FCFS-based scheduling algorithm.

GSA mechanism is for beacon mode to meet the delay constraints of time-sensitive transactions in star topology. GSA is proved to be optimal and work conserving. Different from the earliest deadline first (EDF) scheduling, which results in bursty transmissions of payloads for transactions with delay constraints, GSA smooths out the traffic of a transaction by distributing the GTSs of a transaction over as many beacon intervals as possible while satisfying the time constraint of the transaction. By doing so, GSA reduces the average services to more transactions. This can significantly benefit many time-sensitive applications, where the starting time of the first message and the stability of traffic have great impact on the performance of these applications.

To satisfy two requirements, one is how many GTSs are needed by the payload and the other is how to arrange these GTSs to satisfy the time constraint, GSA is described in the following three steps. First, it checks if all the transactions are schedulable by adding the new transaction to its current transactions. Second, if the transactions are schedulable, then in Step 2 it not only estimates the delay of serving, but also analyzes the relationship between the delay of serving and the number of GTSs allocated to the delay of serving in each interval. Third, based on this relationship, in Step 3

GSA allocates the minimum number of GTSs to the delay of serving in each beacon interval so that the payload can be maximally spread out. To ensure that all GTSs are maximally utilized and the scheduling of GTSs is optimal, GSA adjusts the allocations of GTSs whenever the payload that needs to be transmitted in a CFP changes. These GTSs are evenly spread out over multiple beacon intervals to ensure a smooth traffic flow between the PAN coordinator and sensor nodes.

Each of these discussed approaches has contributed to improve the performance of GTS allocation mechanism in origianl IEEE 802.15.4. The AGA scheme uses the idea of assigning priorities in a dynamic fashion based on recent GTS usage feedbacks with the consideration of low latency and fairness. And the *i*-Game has the advantage of accepting multiple flows sharing the same GTS, while still meeting their delay requirements. Besides, the knapsack algorithm, which is based on the solution of the knapsack problem, ensures that the radio bandwidth in the GTS is utilized in an optimal manner. Furthermore, the GSA mechanism smoothes out the traffic of a transaction by distributing the GTSs of a transaction over as many beacon intervals as possible while satisfying the time constraint of the transaction. By doing so, GSA reduces the average services to more transactions.

Nevertheless, they also have certain drawbacks to some extent. In *i*-Game and GSA, for example, the information of delay requirements needs to be exchanged with the controller, which incurs signaling overhead. The GSA scheme also has high computational complexity due to the execution of a number of algorithms. In the *i*-Game approach, since the algorithm starts the GTS allocation from the last time slot in a round-robin manner, it may fail to serve a flow with hard real-time deadline, which needs to be assigned the first GTS in the CFP. Additionally, it requires a control packet for flow specification in the higher layer. The knapsack algorithm does not provide a detailed priority differentiation mechanism and AGA scheme also has implementation overhead since extra information for devices shall be recorded to allocate GTS resources. Furthermore, energy consumption issue should also be a major concern. All of these above limitations require our future research. In spite of the difficulty of developing an appropriate approach meeting all requirements, we shall do the best to cater specifical needs in different conditions.

5. Cross-Period Approaches

In a cross-period approach, the length of CAP or CFP is dynamically adjusted to various operating conditions. And usually, such changes may have significant impact on both CAP and CFP performance of IEEE 802.15.4 protocol. Hence, setting BO and SO has become one of the most important tasks of the PAN coordinator to determine the superframe structure. Koubaa et al. in [29], analyzed the impact of BO and SO on the performance of slotted CSMA/CA and showed that higher superframe orders provide better network throughput than lower superframe orders due to their increased immunity against the CCA deference symptom.

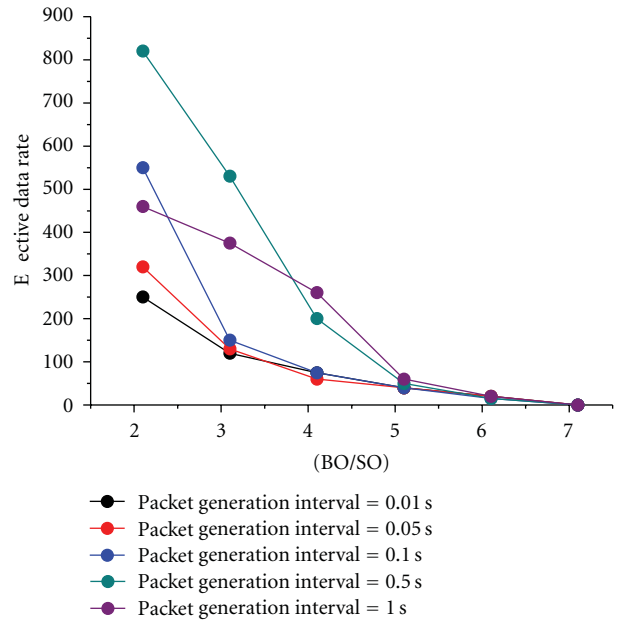


Figure 9: Effective data rate.

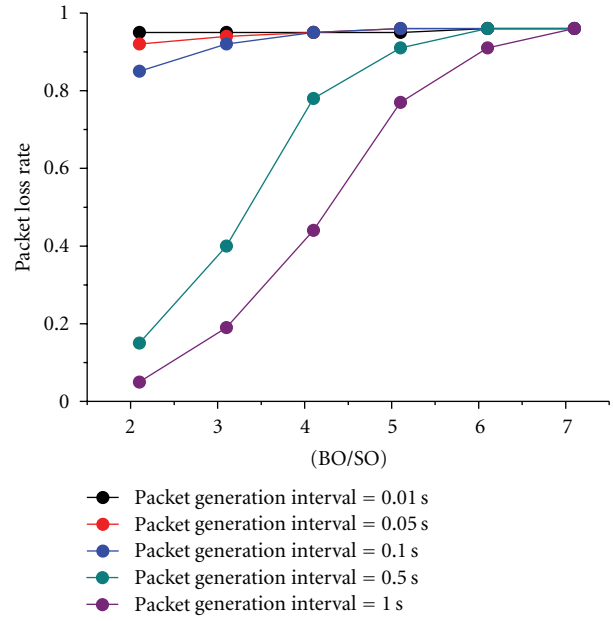


Figure 10: Packet loss rate.

A simulation study of effective data rate and packet loss rate has been provided in [30]. Figure 9 shows the measured effective data rate (SO = 1). We can find that as the value of BO decreases, effective data rate grows gradually. This is mainly because the smaller BO resulting in higher duty cycle can achieve larger bandwidth, which implies larger effective data rates. Figure 10 gives the measured packet loss rate (SO = 1). It has been shown that for the same packet-generation interval, a higher BO leads to a smaller packet loss rate. This is because under the same traffic load, the

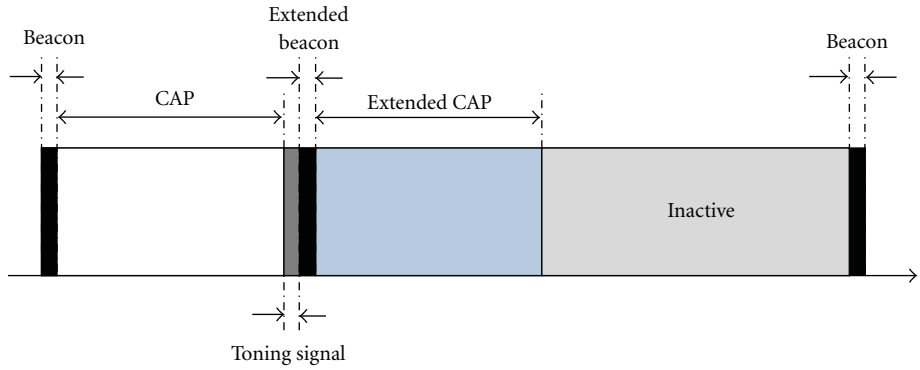


Figure 11: Superframe structure of IEEE 802.15.4 with an extended contention access period.

smaller BO resulting in larger duty cycle enables the network to transmit more packets.

Jeon et al. illustrated PECAP (priority-based delay alleviation algorithm) in [31] about how to set BO and SO at the end of the CAP. In this algorithm, the active period is temporally increased to reduce the sleep delay. Nodes having high-priority packets will request the coordinator to execute an extended CAP by sending a priority toning signal. Thus nodes that have high-priority packets can alleviate delay due to the less contentious environment. Figure 11 shows the superframe structure when the PECAP algorithm is applied. The key advantage of the PECAP algorithm is that it provides exclusive transmission opportunities to the high-priority packets and transmissions of important data can be ensured with timeliness guarantees.

Another example of cross-period approaches is the AGA scheme. Huang et al. proposed a threshold value T_h, which is dynamically adjusted and depends partly on the BO value, due to the consideration of CAP and CFP traffic load. When the CFP traffic load is light, GTS resources are transferred for contention-based access in CAP to filter unnecessary GTS allocation. Besides, the AGA scheme takes advantage of BO changes flexibly. As the BO increases, there is a higher probability that many devices have requested GTS service in the superframe. Hence, in such cases, a more strict threshold value is set to prevent the scarce GTS resources from distributing to those devices with extremely low priorities.

6. Conclusions

In this paper, we mainly focus on limitations of original IEEE 802.15.4 MAC specifications in contention access period, contention-free period, and overall cross-period, respectively. A variety of adaptive and real-time protocols, which are to overcome these constraints, have also been presented and discussed. While it is true that existing research works have significant impact on improving network performance, in terms of adaptive and real-time guarantees, a great number of problems still exist, such as high latency, great energy consumption, system complexity, and implementation overhead. Requirements of all aspects usually cannot be satisfied simultaneously. Hence, our further study on IEEE 802.15.4 MAC is badly needed for developing a more comprehensive and appropriate protocol catering various needs in real operating conditions.

Acknowledgments

This paper is partially supported by Nature Science Foundation of China under Grant no. 60903153, the Fundamental Research Funds for Central Universities (DUT10ZD110), the SRF for ROCS, SEM, and DUT Graduate School (JP201006).

References

[1] F. Shu, N. H. Malka, and W. Chen, "Building automation systems using wireless sensor networks: radio characteristics and energy efficient communication protocols," *Electronic Journal of Structural Engineering*, pp. 66–73, 2009.

[2] Z. Y. Yao, Y. L. Sun, and N. H. El-Farra, "Resource-aware scheduled control of distributed process systems over wireless sensor networks," in *Proceedings of the American Control Conference (ACC '10)*, pp. 4121–4126, July 2010.

[3] M. Marin-Perianu, C. Lombriser, O. Amft, P. Havinga, and G. Tröster, "Distributed activity recognition with fuzzy-enabled wireless sensor networks," *Lecture Notes in Computer Science*, vol. 5067, pp. 296–313, 2008.

[4] B. Aygün and V. C. Gungor, "Wireless sensor networks for structure health monitoring: recent advances and future research directions," *Sensor Review*, vol. 31, no. 3, pp. 261–276, 2011.

[5] C. Emmanouilidis and P. Pistofidis, "Machinery self-awareness with wireless sensor networks: a means to sustainable operation," in *Proceedings of the 2nd Workshop on Maintenance for Sustainable Manufacturing*, pp. 43–50, Verona, Italy, May 2010.

[6] Y. M. Chen, W. Shen, H. W. Huo, and Y. Z. Xu, "A smart gateway for health care system using wireless sensor network," in *Proceedings of the 4th International Conference on Sensor Technologies and Applications*, Verona, Italy, July 2010.

[7] IEEE Standard for Information Technology Part 15.4, Wireless Medium Access Control Layer (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (LR-WPANS), IEEE Standard 802.15.4 Working Group Std., 8 September 2006.

[8] ZigBee Alliance, "The ZigBee Specification version 1.0, 2007".

[9] I. Ramachandran, A. K. Das, and S. Roy, "Analysis of the contention access period of IEEE 802.15.4 MAC," *ACM Transactions on Sensor Networks*, vol. 3, no. 1, 2007.

[10] C. Cano, B. Bellalta, J. Barceló, and A. Sfairopoulou, "A novel mAC protocol for event-based wireless sensor networks: improving the collective QoS," *Lecture Notes in Computer Science*, vol. 5546, pp. 1–12, 2009.

[11] S. Nandi and A. Yadav, "Adaptation of MAC layer for QoS in WSN," *Trends in Network and Communications*, vol. 197, pp. 1–10, 2011.

[12] G. W. Lee, J. H. Lee, S. J. Lee, and E. N. Huh, "An efficient analysis for reliable data transmission in wireless sensor network," in *Proceedings of the IEEE Asia-Pacific Services Computing Conference (APSCC '10)*, 2010.

[13] A. Koubaa, M. Alves, B. Nefzi, and Y. Q. Song, "Improving the IEEE 802.15.4 slotted CSMA/CA MAC for time-critical events in wireless sensor networks," in *Proceedings of the Workshop on Real Time Networks (RTN '06)*, July 2006.

[14] J. G. Ko, Y. H. Cho, and H. Kim, "Performance evaluation of IEEE 802.15.4 MAC with different backoff ranges in wireless sensor networks," in *Proceedings of the 10th IEEE Singapore International Conference on Communications Systems (ICCS '06)*, October 2006.

[15] Z. J. Chen, C. Lin, H. Wen, and H. Yin, "An analytical model for evaluating IEEE 802.15.4 CSMA/CA protocol in low-rate wireless application," in *Proceedings of the 21st International Conference on Advanced Information Networking and ApplicationsWorkshops (AINAW '07)*, pp. 899–904, May 2007.

[16] C. Y. Leea, H. I. Choa, G. U. Hwanga, Y. Dohb, and N. Parkb, "Performance modeling and analysis of IEEE 802.15.4 slotted CSMA/CA protocol with ACK mode," *International Journal of Electronics and Communications*, vol. 65, no. 2, pp. 123–131, 2011.

[17] B. Gao, C. He, and L. G. Jiang, "Modeling and analysis of IEEE 802.15.4 CSMA/CA with sleep mode enabled," in *Proceedings of the 11th IEEE Singapore International Conference on Communication Systems (ICCS '08)*, pp. 6–11, November 2008.

[18] G. Anastasi, M. Conti, and M. Di Francesco, "A comprehensive analysis of the MAC unreliability problem in IEEE 802.15.4 wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, 2011.

[19] V. P. Rao and D. Marandin, "Adaptive backoff exponent algorithm for Zigbee (IEEE 802.15.4)," *Lecture Notes in Computer Science*, vol. 4003, pp. 501–516, 2006.

[20] T. H. Kim and S. Choi, "Priority-based delay mitigation for event-monitoring IEEE 802.15.4 LR-WPANs," *IEEE Communications Letters*, vol. 10, no. 3, pp. 213–215, 2006.

[21] M. Kim and C. H. Kang, "Priority-based service-differentiation scheme for IEEE 802.15.4 sensor networks in nonsaturation environments," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 7, pp. 3524–3535, 2010.

[22] A. Koubâa, M. Alves, E. Tovar, and A. Cunha, "An implicit GTS allocation mechanism in IEEE 802.15.4 for time-sensitive wireless sensor networks: theory and practice," *Real-Time Systems*, vol. 39, no. 1–3, pp. 169–204, 2008.

[23] Y. K. Huang, A. C. Pang, and H.-N. Hung, "An adaptive GTS allocation scheme for IEEE 802.15.4," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 5, pp. 641–651, 2008.

[24] X. J. Dong, "Adaptive polling algorithm for PCF mode of IEEE 802.11 wireless LANs," *Electronics Letters*, vol. 40, no. 8, pp. 482–483, 2004.

[25] S. Yoo, P. K. Chong, D. Kim et al., "Guaranteeing real-time services for industrial wireless sensor networks with IEEE 802.15.4," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3868–3876, 2010.

[26] M. A. Hanson, H. C. Powell, A. T. Barth et al., "Body area sensor networks: challenges and opportunities," *IEEE Computer*, vol. 42, no. 1, pp. 58–65, 2009.

[27] B. Shrestha, E. Hossain, S. Camorlinga, R. Krishnamoorthy, and D. Niyato, "An optimization-based GTS allocation scheme for IEEE 802.15.4 MAC with application to wireless body-area sensor networks," in *Proceedings of the IEEE International Conference on Communications, ICC 2010*, Cape Town, South Africa, May 2010.

[28] C. Na, Y. Yang, and A. Mishra, "An optimal GTS scheduling algorithm for time-sensitive transactions in IEEE 802.15.4 networks," *Computer Networks*, vol. 52, no. 13, pp. 2543–2557, 2008.

[29] A. Koubaa, M. Alves, and E. Tovar, "A comprehensive simulation study of slotted CSMA/CA for IEEE 802.15.4 Wireless Sensor Networks," in *Proceedings of the 6th IEEE International Workshop on Factory Communication Systems*, pp. 183–192, Torino, Italy, June 2006.

[30] F. Xia, A. Vinel, R. X. Gao, L. Q. Wang, and T. Qiu, "Evaluating IEEE 802.15.4 for cyber-physical systems," *Eurasip Journal on Wireless Communications and Networking*, vol. 2011, Article ID 596397, 14 pages, 2011.

[31] J. Jeon, J. W. Lee, H. S. Kim, and W. H. Kwon, "PECAP: priority-based delay alleviation algorithm for IEEE 802.15.4 beacon-enabled networks," *Wireless Personal Communications*, vol. 43, no. 4, pp. 1625–1631, 2007.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 257101, 12 pages
doi:10.1155/2011/257101

Research Article

A Reliable and Efficient MAC Protocol for Underwater Acoustic Sensor Networks

Junjie Xiong, Michael R. Lyu, and Kam-Wing Ng

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

Correspondence should be addressed to Junjie Xiong, jjxiong@cse.cuhk.edu.hk

Received 14 February 2011; Revised 10 May 2011; Accepted 12 July 2011

Underwater acoustic sensor networks (UWASNs) are playing a key role in ocean applications. Unfortunately, the efficiency of UWASNs is inferior to that of the terrestrial sensor networks (TWSNs). The main reasons are as follows: (1) UWASNs suffer long propagation delay; (2) UWASNs are limited by the narrow bandwidth. Many MAC protocols are proposed to improve the efficiency of UWASNs. However, their improvement is not enough. Moreover, few of them consider the reliability of UWASNs even though the packet loss can fail the applications. Actually, a few of the protocols employ the traditional acknowledgment (ACK) mechanism, but they suffer the throughput degradation a lot. In this paper, first, we propose a protocol called RAS, a priority scheduling approach for multihop topologies. RAS is more efficient in throughput and delay performance. Then, we propose a reliable RAS called RRAS that obtains a tradeoff between the reliability and the efficiency. RRAS designs an ACK and retransmission mechanism, that is, different from the traditional one so that it can maintain a comparable throughput when improving the reliability. Extensive evaluations are conducted to verify that RAS is efficient and RRAS is a tradeoff on reliability and efficiency.

1. Introduction

As a crucial ingredient of cyber physical systems, wireless sensor networks (WSNs) facilitate the interactions between human beings and the physical world through sensing, monitoring, and controlling. Being a type of wireless sensor networks, underwater acoustic sensor networks (UWASNs) [1] draw a lot of interest in ocean applications, such as ocean pollution monitoring, ocean animal surveillance, oceanographic data collection, assisted navigation, and offshore exploration, UWASN is composed of underwater sensors that engage sound to transmit information collected in the ocean. The reason to utilize sound is that radio frequency (RF) signals used by terrestrial sensor networks (TWSNs) can merely transmit a few meters in the water [2].

While UWASNs and TWSNs are similar, there still exist many differences that make UWASNs less efficient than TWSNs [1–3]. First and most importantly, the propagation delay of UWASNs is far larger than that of TWSNs. Acoustic signals propagate at about 1500 m/s underwater, while RF signals travel at the speed of light in the air. To transmit a data packet over 1500 meters, it takes 1 s underwater and 5 μs in the air. Due to the high propagation delay of acoustic signals, the network performance of UWASNs cannot be achieved as highly as that of TWSNs. Second, the sound bandwidth is much narrower than RF bandwidth (e.g., 10 kbps Versus 10 Mbps). Consequently, we should utilize the bandwidth in UWASNs more efficiently. Third, the acoustic sensor is more expensive than the terrestrial sensor, and thus, the sensor deployment in the water is more sparse. The average distance among acoustic sensors is usually several hundred meters.

Many existing methods aim at collision avoidance and improving the efficiency of UWASNs, but they are not reliable. APCAP [2] utilizes the maximum propagation delay to avoid collisions and MAC level pipelining to increase efficiency. To avoid the poor throughput caused by using maximum propagation delay, Peleato and Stojanovic apply a shorter delay to avoid collision when the communicating nodes are close to each other [4]. In addition, two ALOHA-based MAC protocols [5] improve the efficiency by reducing the RTS/CTS handshaking and avoiding collisions with the information from the overheard packets. Although all these MAC protocols help improve the efficiency of UWASNs, none of them consider acknowledgments (ACKs) or retransmissions. Since the packet loss may fail the applications, it is very important to maintain a certain level of reliability with ACK and retransmission mechanisms. For example, when a fire breaks out, such event report packets should

arrive at the BS in time. Since they may be lost due to the volatile environment, they should definitely be equipped with retransmission scheme.

On the other hand, existing methods that include ACK mechanisms are not as efficient as the previous techniques with no ACK mechanisms. For example, slotted FAMA [6], UW-FLASHR [7], and the reservation MAC protocol proposed in [8] realize the traditional ACK technique, but they focus on collision avoidance rather than on ACK and the retransmission effects.

Due to the volatile wireless environment, packet loss is very common in UWASNs [1, 9]. Unlike packet collision that can be reduced or manipulated at a very large extent, this kind of packet loss is out of human control. As a result, we should implement ACK mechanism and the corresponding retransmissions so as to improve the network and application reliability. Considering the innate inferior performance and low bandwidth of UWASNs, we should also avoid deteriorating the efficiency too much by enabling ACK and retransmissions. In this paper, to improve the network efficiency, we first design a protocol called RAS (routing and application-based scheduling protocol), based on which we propose a reliable RAS called RRAS to achieve a tradeoff between the reliability and the efficiency.

RAS enables parallel transmissions and utilizing the information from both the routing and application layer it is an efficient priority scheduling at the MAC layer of the base station (BS) [10]. However, it does not require the global positions of all nodes, because it calculates the rough propagation delays between every node pair in the network through the initial synchronization packet exchange that is proposed by Tracy and Roy [8]. Different from RAS whose period is composed of working portion and sleeping portion, RRAS divides the sleeping portion into a NACK-retransmission portion and the sleeping portion. The NACK-retransmission mechanism of RRAS can improve the network reliability.

We summarize our contributions as follows:

(1) We design an efficient priority scheduling protocol called RAS at the MAC layer of BS.

(2) We propose RRAS to improve the network reliability.

(3) Extensive evaluations are conducted to show that RAS is efficient, and RRAS achieves higher reliability than RAS while achieving comparable throughput performance.

In the remainder of this paper, Section 2 describes related work. As the basis of our new protocol RRAS, the RAS protocol is introduced in Section 3. RRAS protocol is designed in Section 4, and its performance is evaluated in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

Recently, there are extensive research efforts focusing on improving the performance of UWASNs, which are surveyed as follows.

Slotted FAMA [6] is a handshaking-based protocol designed to avoid collisions caused by the hidden terminal problem in UWASNs. It synchronizes all nodes and makes all the transmissions start at the beginning of a slot. The slot length is the sum of the maximum propagation delay and the transmission time of a CTS (clear to send) packet. While it can efficiently avoid collision caused by the long propagation of UWASNs and also enable ACK mechanism, it does not utilize the long propagation delay to improve the throughput and delay performance. Our RAS can achieve higher efficiency with a compact schedule that allows high-level parallelization.

The reservation MAC protocol proposed in [8] avoids collisions and improves the bandwidth utilization by employing two channels: a control channel for RTS/CTS handshake and a data channel for data transmission. It increases the network throughput by dividing a larger group of collision sources into smaller ones. While it realizes traditional ACK technique, it does not investigate the retransmission effects. In addition, it is based on a single-hop topology, and the gateway is in charge of the control packet and data packet coordination. Hence, this MAC protocol does not suit the multihop situation, whose complexity requires a different analysis. In contrast, both RAS and RRAS can be applied in either one-hop or multihop topologies.

APCAP identifies that the traditional CSMA leads to poor performance in UWASNs due to the long propagation delay [2]. Therefore, it utilizes the maximum propagation delay to avoid collisions. To avoid the poor throughput caused by such method, Peleato and Stojanovic reduce the delay used to avoid collision [4]. In addition, APCAP employs MAC level pipelining to increase efficiency. Another MAC protocol UW-FLASHR [7] also implements MAC level pipelining. The difference is that APCAP is based on an adaptive and distributed RTS/CTS handshake while UW-FLASHR employs both TDMA mechanism and RTS/CTS handshaking. APCAP do not consider acknowledgements (ACKs) for the data packets while UW-FLASHR enables traditional ACK mechanisms.

Regarding the RTS/CTS handshaking transmission model as inefficient in UWASNs, Chirdchoo et al. propose two ALOHA-based MAC protocols [5]. This MAC protocol also requires the knowledge of propagation delays between every node pair in the network. Then, it uses the information from the overheard packets to avoid collisions. Its simple and clever mechanisms improve the efficiency, but it does not consider ACK mechanisms, either.

Since packet loss is very common in wireless sensor networks [1, 9], our RRAS can greatly improve the network and application reliability.

3. RAS Protocol

The main purpose of RAS is to design an efficient schedule for all the sensor nodes on when to send and receive DATA packets. First, we introduce the RAS process. Then, we focus on the schedule calculation.

3.1. Overview of RAS Protocol. The typical application we discuss is the ocean bottom surveillance application, in which all nodes generate the same amount of data and send them to the BS periodically with cycle T_c. RAS's practicability is guaranteed by the applications features: (2) the maximum one-way propagation delay of UWASNs is very long, for example, 1000 ms, and thus the synchronization accuracy requirement is low; (3) the networks are sparsely (because of high-cost sensors) and statically deployed.

Before schedule calculation, RAS needs synchronization in the networks. Although there exist many synchronization methods [11, 12], for the sake of simplicity and low cost, RAS only requires coarse synchronization through the initial packet exchange or the information piggybacked in the received packets as The authers in [7, 8] do. Meanwhile, it can calculate the rough propagation delays between nodes by checking time stamps during the synchronization process [8].

With synchronization, the nodes can work and sleep periodically [13]. In Figure 1, the RAS cycle T_c is divided into two portions. One portion is the sleeping period, the other is the working period T_w which is divided into many time slots T_s. One time slot is composed of T_d, the time duration for transmitting data, and T_g, the guard time for avoiding collisions introduced by imprecise synchronization and propagation time calculation. The actual value of T_g is determined by the real deployment environment. If there is a data burst due to abnormal events, nodes can transmit them in the following sleeping period by notifying the related nodes in advance. In this way, data burst does not require the schedule update. In the following, we focus on analyzing the working schedule in one cycle.

It is practical to employ static routing, because the networks are static. In addition, the monitoring applications only require data transmissions from the sensor nodes to the BS. Then, the BS calculates the number of data to be transmitted and received at each node. Finally, the BS can calculate for all the sensor nodes the working schedule on when to send and receive data, and broadcast the schedule to all the sensor nodes to follow for a long time. The steps are shown in Algorithm 1.

These processes cost little efforts, because they do not require frequent updates. Therefore, the major goal is how to make the working period of the whole network as short as possible so as to save the energy and improve the network efficiency, that is, how to design an efficient schedule for a cycle.

3.2. Scheduling Principles. The transceiver cannot receive when it is transmitting, and collision will occur at a node when it receives more than one packet [14]. (This corruption is called interference. Interference packets at a node are divided into two types: the first type is for packets that are not destined for the node but are within the node's communication range *RR*. The other type is for packets that are beyond the node's communication range but are within the interference range *RI*. Usually, the relation [15] between *RR* and *RI* is: $2 \times RR \leq RI \leq 3 \times RR$.) In order to avoid collision, we define the following scheduling principles.

(1) A DR duration must not overlap any DT duration.

(2) A DR duration must not overlap any IR duration.

(3) A DR duration must not overlap any other DR duration.

(4) A DT duration and IR duration(s) can overlap.

Since there is no data from the BS to sensor nodes, we can design a compact schedule by only allocating time for data from sensor nodes to the BS. Thus, the next principle is as follow:

(5) No DR from ith hop node to $(i+1)$th hop node.

The goal of data transactions is to guarantee successful receptions, and we arrive at the last principle.

(6) A node considers DR duration as the scheduling basis rather than DT or IR duration.

With DR as the scheduling basis, we do not have to consider whether IR will overlap other IRs and DTs. In addition, we can increase the throughput and reduce delay by making nodes transmit or receive instead of idling whenever no DR is overlapped.

3.3. Scheduling of RAS Protocol. To accomplish the goal of an efficient schedule, RAS implements MAC level pipelining by interlacing multiple data transmissions in a schedule as long as no collision is triggered. During the interlacing process, RAS assigns the slots to nodes with heavier traffic first that is, nodes with heavier traffic are given higher priority in data transmission and reception, and thus other nodes can only transmit or receive in the time slots that are not taken up by the heavier-traffic nodes. In this way, the traffic load can be better balanced, and the efficiency can be improved as well.

We call the schedule length calculated with RAS as L_1. In RAS, the scheduling element corresponds to one data transaction in which the data transmission and reception will last for several time slots. The scheduling element is composed of one DT duration, one DR duration, and several IR durations.

According to scheduling principle (6), we will schedule all the elements generated in a cycle equals by scheduling all the data receptions in a cycle. Let $\mathbb{S}_i = \{m : \text{node } m \text{ is } i \text{ hops from the BS}, m \in \mathbb{S}\}$. Assuming a node's distance from the BS is proportional to its hop distance to the BS, the following is the priority scheduling steps of RAS algorithm.

Step 1: Schedule the BS's DR from 1-hop nodes.

Step 2: Schedule the DR tier by tier: from inner tier to outer tier, that is, from DR of nodes in $\mathbb{S}_i$ to DR of nodes in $\mathbb{S}_{i+1}$, $i \in \{1, 2, 3, \ldots, H-1\}$, H is the maximum hop distance to the BS.

Step 3: For each node $m \in \mathbb{S}_i$ that is going to receive data packets from its children $C_{mj} \in \mathbb{S}_{i+1}$, $j = 1, 2, \ldots, K_m$, arrange its DR from its children alternatively. For example,

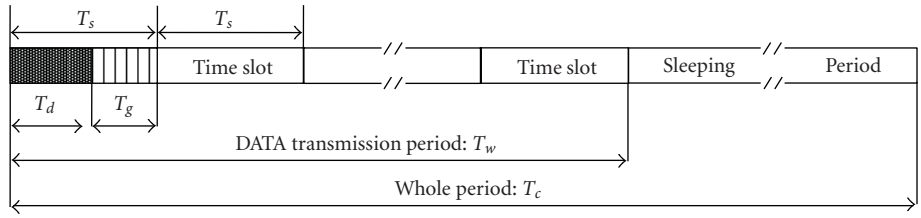


Figure 1: RAS cycle.

(1) Coarse synchronization through the initial packet exchange
(2) Calculate the rough propagation delay between nodes through the previous packet exchange
(3) Calculate static routing
(4) Calculate the number of data to be transmitted and received at each node
(5) CalcSchedule() /*Algorithm 3*/
(6) The BS broadcasts the routing table and the schedule to all its children with high power

Algorithm 1: RAS protocol at the BS (it runs at the initialization phase).

node 1 has two children A and B. In a cycle, each of them send 3 packets P_{Ai}, P_{Bi} to node 1, and $i \in 1, 2, 3$. Then, one possible DR sequence at node 1 is: $P_{A1}, P_{B1}, P_{A2}, P_{B2}, P_{A3}$, and$P_{B3}$.

The reason to schedule the BS's DR first is that the topmost goal is for the BS to receive all the data generated by all other nodes in a cycle. The reason to prioritize the inner tier nodes over the outer tier nodes is that the inner-tier nodes are affording much heavier traffic. It is unfair for them to share the same bandwidth with other light-traffic nodes. In addition, the packets forwarded by the inner tier nodes are forwarded more hops than those forwarded by the outer tier nodes. Colliding or dropping those packets would cost more efforts to retransmit. Finally, to alternate the receptions among the children provides load balancing of the nodes. If fairness is not considered for a parent's children, a few children might drop packets due to congestion, whereas other children's queues are far from full. Therefore, when we alternate the children's transmissions, we improve the fairness. The three steps are shown in Algorithm 2. How RAS works at the sensor nodes is shown in Algorithm 3.

3.4. Analysis of the RAS Scheduling Algorithm. The time used to calculate the schedule of RAS is less than 1 sec in our personal computer, because RAS is a heuristic method with complexity O(N) in which N is the node number.

Since the upper bound for the RAS schedule length (L_1) can be infinitely long, we only discuss L_2, the lower bound for the schedule length. Assuming each node generates P packets in a cycle, then the BS has to receive $N \times P$ packets in total from 1-hop nodes in a N-node network. To receive packets, it has to wait for at least the propagation time T_p of one packet. Therefore, the shortest time for receiving all the $N \times P$ packets is $N \times P \times T_s + T_p$. We call this time L_2 as the *lower bound*, which cannot be achieved in large-scale networks due to interferences.

4. RRAS Protocol

Since packet loss is very common in UWASNs [1, 9], RAS is not reliable. By enlarging the guard time, we can avoid collisions caused by imprecise synchronization. Hence, in our case, we focus on the *packet loss caused by the volatile wireless environment*. Because RAS does not considering the ACK and retransmission problem, we propose a reliable RAS called RRAS. RRAS employs the RAS scheduling to transmit all the data efficiently. For the data packets lost during the scheduling, RRAS utilizes the NACK-retransmission mechanism to improve the overall system reliability.

4.1. Overview of NACK-Retransmission Mechanism. In the scheduling of RAS, each node n knows n_c, the number of packets that should be received from each of its child node c, and n_0, the number of packets generated by itself. The total packets that should be sent to its parent are $n_0 + \sum_c n_c$. In RRAS protocol, if in a cycle, a node does not receive the expected packets, it would keep the packet loss in mind and perform retransmission in the following *retransmission period*.

Retransmission period is part of an RRAS cycle as shown in Figure 2. RRAS cycle is designed from RAS cycle: the data transmission period of RAS is kept, while the sleeping period of RAS is divided into the retransmission period and a shorter sleeping period. In this way, the data lost in data transmission period T_w can be retransmitted in retransmission period T_r. Obviously, RRAS protocol keeps the nodes working for a longer period than RAS for the sake of reliable transmission. However, since the unavoidable packet loss might fail the applications, it is worthy of the extra time T_r to keep all the nodes waiting for possible data retransmission requirement.

The packet loss caused by the volatile wireless environment is random, thus we do not know which packets

```
(1) Parent = BS; hop = 1. //schedule the BS's DR from 1-hop nodes
(2) while hop ≤ maxhop. do
(3)   while Parent has children. do
(4)     while Parent has data to receive from its children. do
(5)       if Parent is idle in the Time Slot Slot. then
(6)         With global information, Parent searches its entire children to alternatively find
            a child whose transmission results in its reception at the Slot.
(7)         if Parent finds a suitable child. then
(8)           schedule the child's transmission and the related reception and interference.
(9)           break searching.
(10)        end if
(11)      end if
(12)      Parent fetches the next Slot for reception.
(13)    end while
(14)    fetch the next Parent to schedule reception.
(15)  end while
(16)  hop = hop + 1. //schedule the DR tier by tier
(17) end while
```

Algorithm 2: CalcSchedule() function at the BS.

```
(1) Node x receives routing information and a schedule from the BS.
(2) Synchronization.
(3) while 1 do
(4)   if node x is in sleeping period. then
(5)     sleep until working period starts.
(6)   end if
(7)   if abnormal events happen. then
(8)     request x's parent to wait for data in sleeping period.
(9)   end if
(10)  if receive data from x's children. then
(11)    receive the data as scheduled.
(12)  end if
(13)  if x has data to send. then
(14)    send the data in working period as scheduled.
(15)  end if
(16) end while
```

Algorithm 3: RAS protocol at the other nodes (it runs at each sensor node after the initialization).

will be lost in the data transmission period T_w, and we cannot schedule the retransmission as we schedule the periodical data transmission. As a result, we employ a NACK-retransmission mechanism, that is, the NACK packet asks for the lost packets, and the retransmission packets reply NACK with the required packets. The detailed explanation follows.

During retransmission period T_r, the node n that has not received the expected packets would send a control packet (NACK) to its child node c whose packets to n are lost during transmission. The NACK contains the packet sequence numbers that n has received from c. On receiving the NACK, c would know which packets to retransmit. If node n failed to receive packets from multiple child nodes, it would send an NACK to the very child nodes one by one. If the retransmissions succeed, no more retry is needed. Otherwise, node n could initiate the corresponding retransmission by sending another NACK.

4.2. Retransmission Mechanism. The more times that node n retry the transmissions, the higher the reliability is. Although such retransmission is aimed at improving the data transmission reliability, it does not promise 100% success which would cost a lot of extra resources. Hence, in our case, we only focus on one-time retry rather on the frequently used three-time retry [13]. If we want to retry a second time, we can simply add another retransmission period to retransmit the data packets lost during the first retransmission period.

After receiving NACK, node c would retransmit the lost data packets in a batch or in a burst to node n. Since the control frame exchanges deteriorate the UWASN efficiency greatly, batch data transmission improves the retransmission throughput and delay performance by reducing the exchanges.

As for the collision avoidance, in order not to degrade the efficiency, we do not use the maximum propagation delay

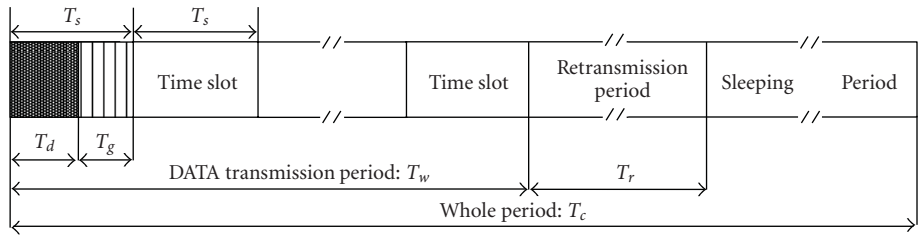


Figure 2: RRAS cycle.

to avoid collisions as [2, 6] do. In addition, carrier sense and RTS/CTS handshaking are efficient in a network where the propagation delay is negligible, but they are inefficient in UWASNs [16]. As a result, we adopt a simple ALOHA mechanism [5, 17]. According to [10], the propagation time is usually larger than the data duration time (100 ms) as long as the distance between the nodes is larger than 150 m. In this situation, even if the two nodes send packets to each other simultaneously, no collision happens. Furthermore, if node A and node B start to send a packet to C at the same time, as long as the difference between A–C distance and B–C distance is larger than 150 m, there will be no collision at node C. For these reasons, we employ simple ALOHA that is, a node could transmit a packet when it is not receiving or transmitting. Although this mechanism does not guarantee collision-free situations, it is more efficient. In case collision happens, further actions can be taken: retransmit the collided packets again or ignore it and accept the current reliability level. In our work, we study the effects of ALOHA through experiments and prove its influence on the efficiency and reliability.

We define the states of a node after the data transmission period as (α, β). The α state is determined by the node itself, while the β state is determined by its parents

$$\alpha = \begin{cases} \text{Y}, & \text{packets from its child nodes are lost,} \\ \text{N}, & \text{packets from it child nodes are not lost,} \end{cases}$$
$$\beta = \begin{cases} \text{Y}, & \text{packets to its parent nodes are lost,} \\ \text{N}, & \text{packets to its parent nodes are not lost.} \end{cases} \quad (1)$$

A node might be in 4 situations: (N,N), (N,Y), (Y,N), and (Y,Y). If it it in state (N,N) with no packet loss, it is free from sending or receiving packets. If it is in state (N,Y), it does not know that its data transmissions to its parents are lost until it receives the NACKs from its parents. In this case, it will only retransmit the data packets designated in the NACKs.

If it is in state (Y,N), it knows that data transmissions from its child nodes are lost. It is also aware that since it fails to forward the lost packets to its parent nodes, its parent nodes will require them through NACK. It should first ask its child nodes for the lost packets by sending NACKs to them. Next, after collecting the lost packets from its child nodes, it should forward them to its parent nodes. Since it does not know the *beta* state, it does not know whether its parent nodes will ask for lost packets generated by its child nodes or packets generated by itself. As a result, even if it has finished collecting the lost packets from its child nodes, it still should wait to retransmit until receiving the NACKs from its parents.

The actions in state (Y,Y) is similar to those in state (Y,N). The only difference is that the node should send the lost packets from both its child nodes and from itself to its parent nodes. In summary, the retransmission will be triggered only after receiving the NACK requirements.

4.3. Retransmission Time. In addition, the length of T_r and extra energy consumption E for reliable transmission is closely related to the packet loss rate R. The overall energy consumption is dominated by the transmit power E_t, as compared to receive power E_r and idle power E_i [18, 19]. If R is low, then E is low. For example, if R is 0, that is, there is no packet loss during T_w, then no retransmission is needed, and all the nodes are idling during T_r with low extra power consumption. If the packet loss rate is very high during T_w, then the retransmission period T_r should be long enough to finish all retransmissions, and the extra power would be a lot. In other words, high packet loss means that the wireless environment is very severe, and retransmission is required no matter what kind of mechanisms are employed for data transmissions.

Given packet loss rate R, to ensure that the time duration is long enough for retransmitting the lost packets, the length of T_r for a given network topology is calculated as follows.

To make sure that T_r can accommodate all retransmission situations, we analyze the worst case. Consider the nodes in the longest route to the BS, from the leaf node a to the BS, and the maximum hop distance is H.

If a DATA transmission from node a to its parent b is lost, then all the nodes in the route need to retransmit the lost data to the BS. A node will retransmit only on receiving the NACK requirement, then a retransmission between a pair of nodes would take at most $2 * (T_p + T_s)$, in which T_p is the maximum propagation delay and T_s is the time slot. The longest time taken for retransmitting the DATA to the BS is consequently

$$2 * H * \left(T_p + T_s\right). \quad (2)$$

The ALOHA retransmission mechanism only restricts the order between the NACK and its corresponding retransmission that is, it allows parallel transmission among NACKs

from different nodes. As a result, the time used is much less that is,

$$\left(T_p + T_s\right) + H * \left(T_p + T_s\right). \quad (3)$$

The first term $(T_p + T_s)$ is used to transmit the NACKs, and the second term $H * (T_p + T_s)$ is used to transmit the lost DATA to the BS.

We then consider the case that more than one packet are lost in such a route. The worst situation is that all the packets are lost on the way from a to b. The batch DATA transmission mode reduces the NACK-retransmission handshake, and keeps the handshake round to be 1 among a node pair. Hence the longest time taken for retransmitting all the DATA to the BS is

$$\left(T_p + T_s\right) + H * \left(T_p + L_p * T_s\right), \quad (4)$$

in which L_p is the total number of DATA packets lost,

$$L_p = A_p * R, \quad (5)$$

given A_p, the number of all the packets to be transmitted in a cycle.

Next, consider the case that more than one route, r_o routes, experience data loss. The worst situation is that all the routes are H hops from the BS, and there is no overlap between any of the routes until they converge at the BS. In this situation, the batch DATA transmission mechanism cannot be applied to reduce the time. Then, we get the length of T_r, that is, long enough for most retransmission situations as:

$$T_r = r_o * \left(T_p + T_s\right) + r_o * H * \left(T_p + L_p * T_s\right). \quad (6)$$

When the rough distances between nodes and the routing are known, H and maximum r_o are determined, then T_r can be calculated. However, this value is too large for a real deployment and hence wastes retransmission time. In fact, not all data losses happen on the way from the farthest nodes to their parents. In addition, the interference between the routes may be very light, and the transmissions in each route can happen simultaneously without affecting each other. Furthermore, the propagation time to transmit 1 DATA packet for 1 hop is less than the maximum propagation delay T_p. As a result, the actual retransmission time needed is far less than that indicated in (6), and it should be adjusted accordingly. As our later experiments prove, simulation can help us select a better value for T_r.

5. Performance Evaluation

The protocol performance is simulated using the parameters shown in Table 1. We use the ns2 setdest tool to generate networks with size ranging from 9 nodes to 64 nodes [20]. For example, the 64-node network covers a 5 km by 5 km area, and it is connected without holes. The maximum one-way propagation time is 1000 ms calculated from the 1500 m transmission range and sound speed.

Table 1: Parameters for data transmissions.

Parameter	Value
Data rate	10 kbps
DATA packet size	100 bytes
DATA packet duration	80 ms
Sound speed	1500 m/s
Transmission range (communication range)	1500 m
Interference range	3500 m
Guard time	20 ms
Wireless model	TwoRayGround

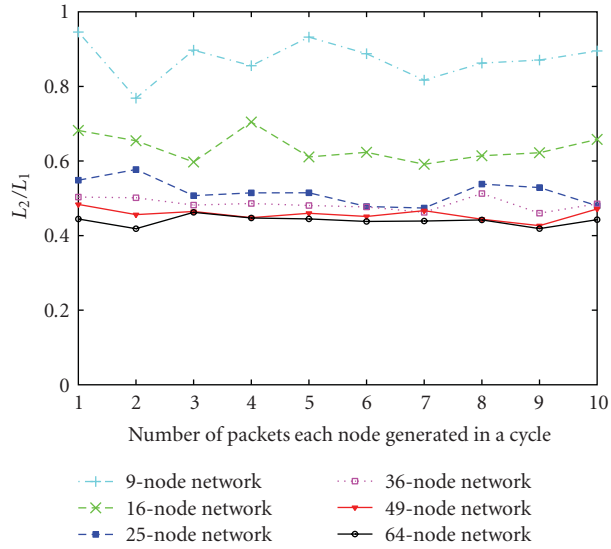


Figure 3: Schedule ratio.

The efficiency performance of RAS is compared with UW-FLASHR [7]. The efficiency and reliability performance of RRAS is compared to RAS and the traditional CSMA/CA used in TWSNs. UW-FLASHR uses control frame handshaking to reserve parallel transmissions. RAS does not employ ACK mechanism, and the traditional CSMA/CA does not adjust the collision avoidance method for UWASNs.

5.1. Schedule Length. Since RRAS uses the schedule of RAS, this section is about RRAS and RAS schedule length. We calculate the lower bound schedule length L_2 and the RAS schedule length L_1 when the number of packets generated by each node in a cycle varies from 1 to 10. Figure 3 shows the value variations of L_2/L_1. The ratio of L_2 to L_1 for each network almost stabilizes at a constant value. Since $L_2 = N \times P \times T_s + T_p$, then L_2 also increases linearly with the network size N. This means that RAS is scalable in calculating the schedule no matter the traffic rate is low or high. Moreover, the ratio of small size networks is higher. The reason is that the hop distances of small-scale networks are 1-hop or 2-hop, and they suffer less from the interferences caused by neighboring nodes. When there are few interferences, the schedule length is reduced.

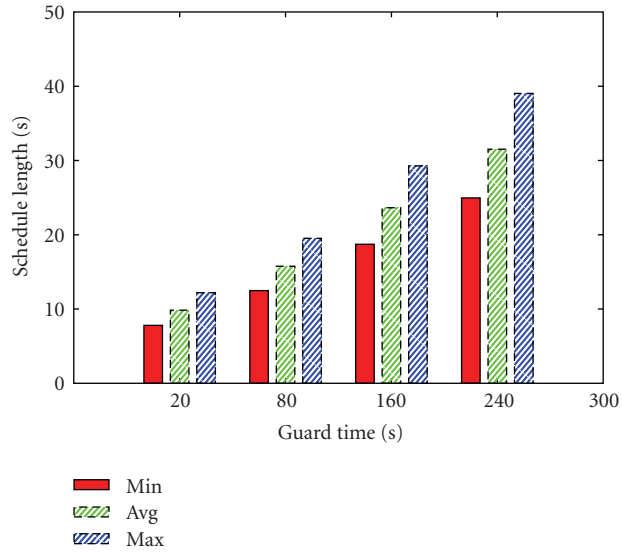


Figure 4: Schedule length.

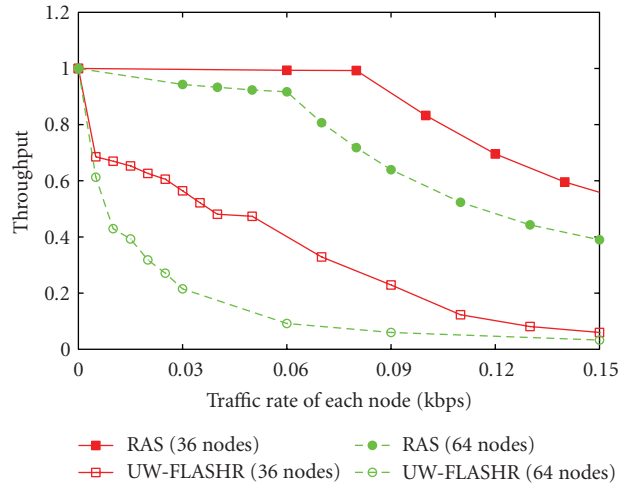


Figure 5: Throughput.

Basically, the larger the guard time, the longer the schedule length. Longer guard time allows more imprecise synchronization while reduces the network efficiency. Figure 4 shows the relation between the guard time and the schedule length for ten 64-node networks. In accordance with our expectation, the schedule length increases linearly with the guard time ranging from 25% of the data duration time (80 ms) to 300% of the data duration time. Therefore, if we deploy a network with high clock drift, we should set the guard time to a longer value. Otherwise, we can use a smaller guard time.

5.2. Throughput of RAS. Due to RAS's high scalability in schedule length demonstrated in the previous subsection, we use the schedule calculated for the case when only one packet is generated in a cycle since this subsection.

To compare RAS with UW-FLASHR [7], an existing MAC protocol designed for high channel utilization, we employ their throughput definition. Throughput is defined by measuring the total number of the intended data packets received by the BS by the total number of data packets generated by all the nodes in a period. Obviously, if the traffic generated at each node is so heavy that it exceeds the maximum capacity of the network, then the throughput would drop and even approaching to 0. Conversely, if the traffic is light, then it is likely that all the data generated will be received by the BS; therefore, the throughput is 1 when there is no traffic.

Although we simulated networks of different sizes, for the sake of conciseness, Figure 5 only compares the throughput of RAS and UW-FLASHR in 36-node networks and 64-node networks. As the traffic rate increases, the throughput of all the networks drops from 1. In addition, 36-node networks are able to afford a much heavier traffic rate than the 64-node networks, because networks with a larger size suffer higher total traffic. Moreover, we notice that the throughput for 36- and 64-node networks with UW-FLASHR dramatically drops from 1 when the traffic rate is not 0. This is because UW-FLASHR performs the slot requirement among the neighbors, and the hidden terminal problem leads to some slot establishment which might cause collisions. On the other hand, RAS performs scheduling based on all nodes' position information, thus no collision happens. Finally, for UW-FLASHR, the heaviest traffic rate these networks could afford is much less than that of RAS. This is because RAS generates a much compacter schedule than UW-FLASHR. RAS arranges the exact time needed by the transmission and reception of each node, while UW-FLASHR reserves the time slots for transmission randomly.

5.3. Average End-To-End Delay of RAS. The end-to-end delay is the period from the time a packet is generated by a node until the time it is received by the BS. Figure 6 shows that the average end-to-end delay increases when the traffic rate increases. Specifically, when the traffic rate is heavy enough to cause congestion in the networks, there are sudden jumps of delay as observed in the figure. By observing the sudden jumps in delay, we find that RAS networks can afford about 4 times higher traffic load than UW-FLASHR networks without collision. Because the scale of 36-node networks is smaller than 64-node networks, their end-to-end delay is also shorter. In addition, when heavily congested, the delay of RAS networks and UW-FLASHR networks stops increasing with traffic rate. The reason is that both RAS and UW-FLASHR are based on TDMA to reserve the channel rather than on CSMA/CA to compete the channel, thus the delay reaches an upper bound. However, the delay upper bound of RAS networks is higher than that of UW-FLASHR networks, this is due to the fact that: the priority scheduling makes the queue utilization of RAS networks higher than that of UW-FLASHR networks (this will be explained in the following subsection). In UW-FLASHR networks, the queue utilization is very low. Most packets from faraway nodes cannot arrive at the BS before being dropped by the heavy-loaded forwarding

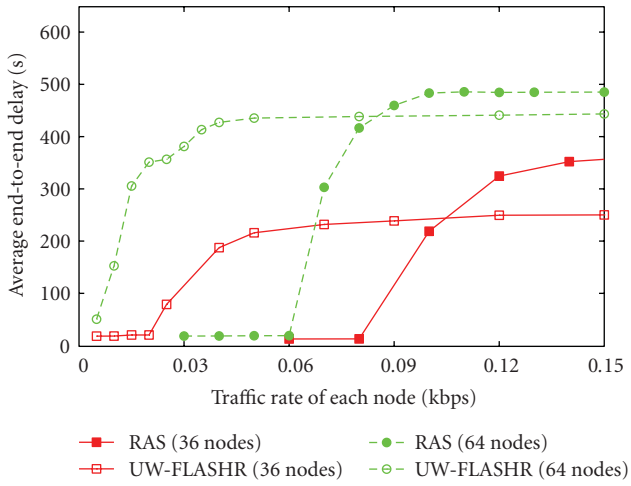


Figure 6: End-to-end delay.

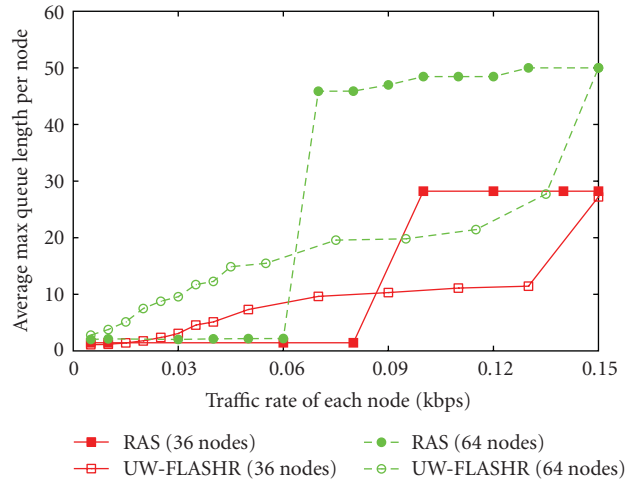


Figure 7: Queue length.

nodes. In other words, most of packets that arrive at the BS are generated by nearby nodes, thus the delay upper bound is lower. In contrast, this phenomenon is alleviated by the priority scheduling in RAS networks.

5.4. Average Maximum Queue Length per Node of RAS. In this subsection, we demonstrate the advantage of RAS in fairness by showing that its queue utilization is fairer than that UW-FLASHR. The queue size of each node is set to 50 in the simulations. If a queue is filled with 50 packets, then further packet arrival will cause one packet to be dropped.

The sudden jump in the queue length is caused by congestion. For example, for the 64-node RAS network, after the traffic rate reaches 0.06 kpbs, the network starts to congest, and the congested packets are put into the queue. If the traffic rate continue to be 0.06 kpbs or higher, more packets will be enqueued until the queue is filled. Since our simulation runs a long time enough at each traffic rate, the queue is full almost all the time when the traffic rate is higher than 0.06 kpbs.

UW-FLASHR does not take the application direction into consideration, nor does it arrange longer time for nodes with heavier traffic. As a result, nodes with heavier traffic (i.e., nodes that are nearer to the BS) would easily accumulate a long queue of packets and suffer queue overflow very soon while nodes with lighter traffic maintain an empty queue. The queue utilization of the nodes in UW-FLASHR is unfair and low. Actually, nodes with heavier traffic experience a larger packet arrival rate, and they need more time to handle the packets. RAS gives higher priority to nodes with heavier traffic by allocating more data transmission time to them, thus their packet leaving rate is also higher. Likewise, nodes with lighter traffic are allotted less time. As a result, the queues of all the nodes are balanced, and the queue utilization is fairer.

Due to similar phenomenon of 36-node networks and 64-node networks, we mainly discuss the case for 64-node networks in Figure 7. When the traffic rate is between 0 kbps and 0.06 kbps, the queue length of RAS networks is shorter than that of UW-FLASHR networks. Due to RAS's capability of affording higher traffic rate, most of the nodes in RAS networks do not have to queue under those traffic rates. Whereas the nodes in UW-FLASHR networks are queueing more and more packets when traffic gets heavy. In addition, RAS networks undergo a jump in queue length when the traffic rate increases from 0.06 kbps to 0.07 kbps. This is because the queue utilization of RAS networks is fairer and higher than UW-FLASHR networks. At traffic rate 0.06 kbps, most of the nodes in RAS networks start to congest. If the traffic rate continues to be 0.06 kpbs or higher, as time goes by, more packets will be enqueued until the queues are full. Our simulation runs a long time enough at each traffic rate, the queue is full almost all the time when the traffic rate is higher than 0.06 kpbs. In contrast, in UW-FLASHR networks, the queues of the few nodes that are next to the BS are congested at a very low traffic rate, while the queues of faraway nodes are empty. Other nodes get congested gradually with the increasing traffic rate, thus the queue length does not surge.

Furthermore, because small-scale networks are less likely to suffer congestion, the queue length of 36-node RAS networks soars at about traffic rate 0.09 kbps rather than at 0.07 kbps. It stabilizes at around 27 rather than at 50 when the traffic rate is larger than 0.11 kbps. Eventually, it will stabilize at 50 when the traffic rate is very high. Nevertheless, 0.15 kbps is not a very high traffic rate for 36-node networks, thus only a majority of the nodes are congested while the others sustain empty queue. Still the queue length of RAS networks is much larger than that of UW-FLASHR networks when congestion happens in RAS networks, which again indicates that the queue utilization of RAS networks is fairer and higher.

In summary, greater maximum queue length allows fairer and higher utilization of the queue. Correspondingly, the delay upper bound of the RAS networks is higher than the UW-FLASHR networks because more faraway packets are capable of arriving at the BS with no overflow in queue.

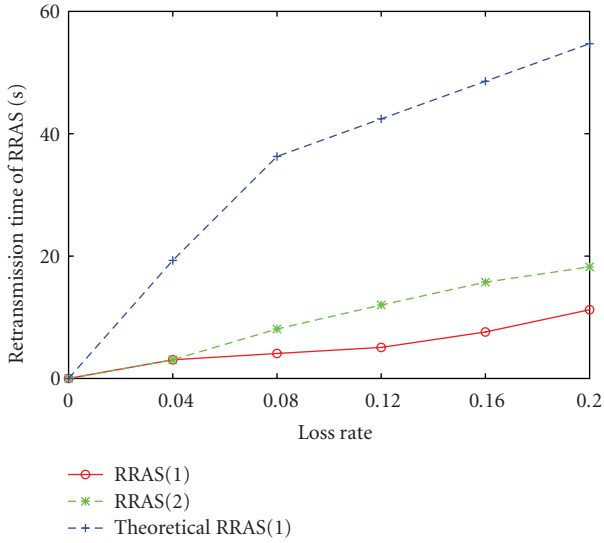


Figure 8: Retransmission time.

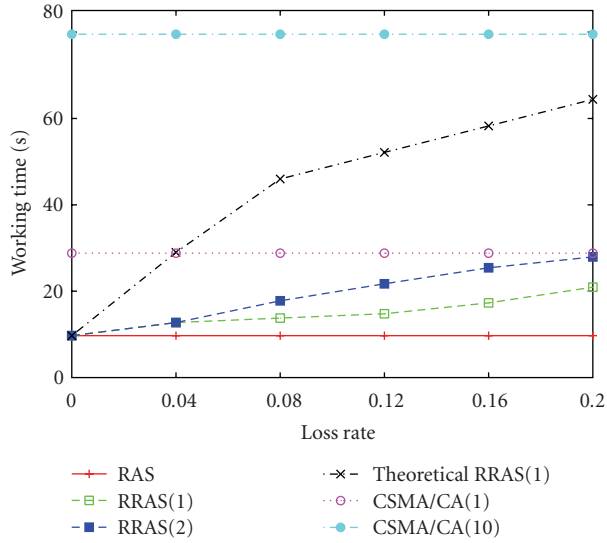


Figure 9: Working time.

5.5. Retransmission Time of RRAS. In this subsection and the following subsections, each node generates 1 packet for the BS periodically. Experiments are performed to determine the actual length of the retransmission period. Although we compare UW-FLASHR with RAS, we cannot compare it with RRAS, because it does not contain retransmission mechanism. According to (6), the theoretical length of the retransmission period is conservatively long, attempting to fit in all possible retransmission communications. Actually, the retransmission time can be much shorter. In Figure 8, RRAS(1) and RRAS(2) are the retransmission time when the retry time is 1 and 2 respectively. Theoretical RRAS(1) is the retransmission time calculated with (5) when the retry time is 1.

In Figure 8, with loss rate (loss is caused by the wireless environment) increasing from 0, the retransmission time of all the three cases increases from 0. RRAS(2) is about a double of RRAS(1), because it retransmits the lost packet one more time. When the loss rate is very low, say 0.04, RRAS(1) and RRAS(2) are equal, because the DATA packets that need retransmission are so few that they can be sent simultaneously. Theoretical RRAS(1) is the longest, because it is an overestimated value that attempts to accommodate all possible retransmission cases. As a result, to save the time spent on retransmission, we can adjust the retransmission time according to the simulation results and the loss rate in the deployment.

5.6. Working Time of RRAS and RAS. Working time is the total time that a node is turned on. During the working time, the node can be sending, receiving, or idling, but not sleeping. For RAS, the working time is the schedule time, and the sensor nodes sleep for the rest of the cycle time. For RRAS, the working time is the schedule length plus the retransmission time. For the traditional CSMA/CA used in TWSNs, the working time is the time used for transmitting all DATA packets to the BS without exceeding the required retry times. Besides RAS, RRAS(1), RRAS(2), and theoretical RRAS(1), Figure 9 also shows the working time of CSMA/CA(1) and CSMA/CA(10), with 1 and 10 as the retry limit, respectively.

Because the traditional CSMA/CA employs RTS/CTS handshaking and does not design the collision avoidance mechanism suitable for UWASNs, many packets are collided and its efficiency is very low. The CSMA/CA has to retransmit the collided DATA packets, and it cannot successfully retransmit all the collided DATA packets to the BS when the retry time limit is 10, let alone when the retry time is 1. As a result, no matter what the loss rate is, the working time of CSMA/CA is a constant. In other words, the protocol for UWASNs should first settle down the negative effects caused by improper collision avoidance methods of the traditional CSMA/CA, and then recovers the packet loss caused by the volatile environment. Figure 9 also demonstrates that (1) the RAS work time is the lower bound of all because it does not retransmit, and, (2), except for CSMA/CA, the working time increases with the increase of retry time and loss rate.

5.7. Success Rate of RRAS and RAS. The success rate is defined as the DATA packets received by the BS divided by the total DATA packets for the BS. Because the RAS does not perform ACK or retransmission mechanism, its DATA success rate is reduced by the loss rate. RRAS can achieve a higher success rate. Figure 10 verifies that compared with RAS, RRAS attains a higher success rate, especially when the loss rate is high. In addition, success rate of RRAS(2) is higher than that of RRAS(1), because RRAS(2) retries one more time. However, according to Figure 8 and Figure 9, one more retransmission of RRAS(2) results longer retransmission and working time. Finally, as discussed previously, the efficiency of CSMA/CA is very low, including the success rate.

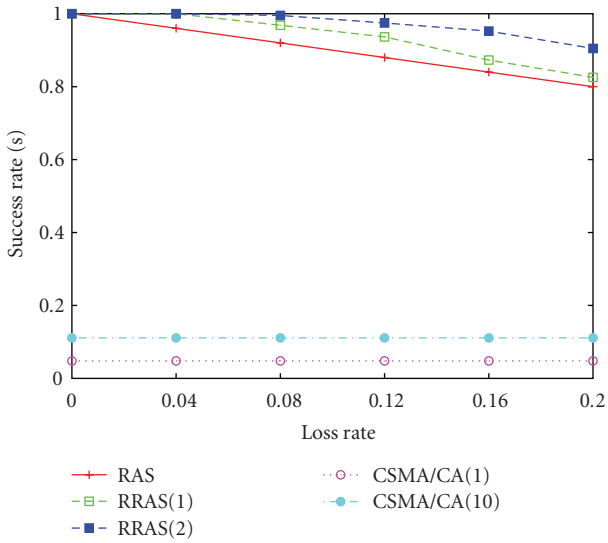


Figure 10: DATA transmission success rate.

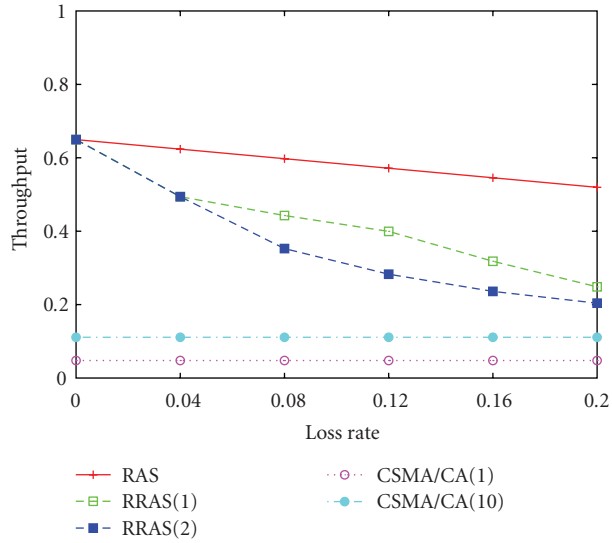


Figure 11: Throughput.

5.8. Throughput of RRAS and RAS. According to [1], the throughput is defined as a function of the sensor node working time and the total DATA packets received by the BS:

$$\text{throughput} = \frac{\text{total DATA packets received by the BS} * T_d}{\text{working time}}. \tag{7}$$

In consistency with the poor success rate, the traditional CSMA/CA efficiency is also very low in Figure 11. As discussed in subsection retransmission time, the throughput of the traditional CSMA/CA is determined by its poor collision avoidance method rather than by the packet loss rate caused by the environment hence, it is flat. According to Figure 11, the throughput of RAS is the highest, but the throughput performance of RRAS is not greatly reduced, especially when the loss rate is not very high. Since the success rate of RRAS is much higher than RAS according to Figure 10, we can observe that RRAS obtains a good tradeoff between the reliability and the throughput.

6. Conclusions

In this paper, we propose an efficient MAC protocol called RAS to improve the efficiency in UWASNs. Then RRAS, the reliable RAS, is implemented to achieve a tradeoff between the reliability and efficiency performance in UWASNs. RAS employs priority scheduling with coarse synchronization and information of the rough propagation delays between each node pair. With these condition and parallel transmissions, RAS can calculate a compact schedule at the BS when static routing is applied and the DATA direction is only from the sensor nodes to the BS. To improve the throughput, RRAS employs the RAS scheduling to transmit the majority of the DATA generated in a cycle to the BS. Then, it designs a NACK-retransmission mechanism to retransmit the DATA packets lost during the previous DATA transmissions. Because this new NACK-retransmission mechanism is based on the RAS scheduling that allows each node to know the packets it should receive during a cycle, it can trigger the NACK requirement after a node identifies its lost packets. Both the NACK and retransmission packet transmissions are distributed. They are based on ALOHA so as to reduce the control frame handshaking and improve throughput. The simulation results demonstrate that RAS is not only efficient, but also obtains good fairness performance. RRAS not only effectively improves the reliability through retransmission, but also attains a comparable throughput. In the future, we are interested in researching the performance of the current MAC protocols when the loss rate is higher than 20% and design a reliable protocol for such situation.

Acknowledgments

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. CUHK4154/10E) and sponsored in part by the National Basic Research Program of China (973) under Grant no. 2011CB302600.

References

[1] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges," *Ad Hoc Networks*, vol. 3, no. 3, pp. 257–279, 2005.

[2] X. Guo, M. R. Frater, and M. J. Ryan, "An adaptive propagation-delay-tolerant MAC protocol for underwater acoustic sensor networks," in *Proceedings of the Oceans*, pp. 1–5, June 2007.

[3] J. Partan, J. Kurose, and B. N. Levine, "A survey of practical issues in underwater networks," in *Proceedings of the 1st ACM International Workshop on Underwater Networks (WUWNet '06)*, pp. 17–24, September 2006.

[4] B. Peleato and M. Stojanovic, "A MAC protocol for ad-hoc underwater acoustic sensor networks," in *Proceedings of the 1st ACM International Workshop on Underwater Networks (WUWNet '06)*, pp. 113–115, September 2006.

[5] N. Chirdchoo, W. S. Soh, and K. C. Chua, "Aloha-based MAC protocols with collision avoidance for underwater acoustic networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 2271–2275, June 2007.

[6] M. Molins and M. Stojanovic, "Slotted FAMA: a MAC protocol for underwater acoustic networks," in *Proceedings of the Oceans Conference*, May 2006.

[7] J. Yackoski and C. C. Shen, "Achieving high channel utilization in a time-based acoustic MAC protocol," in *Proceedings of the 3rd ACM International Workshop on Underwater Networks,(WUWNet '08)*, pp. 59–66, September 2008.

[8] L. T. Tracy and S. Roy, "Short paper: a reservation MAC protocol for ad-hoc underwater acoustic sensor networks," in *Proceedings of the 3rd International Workshop on Underwater Networks, (WUWNet '08)*, pp. 95–98, September 2008.

[9] J. Heidemann, W. Ye, J. Wills, A. Syed, and Y. Li, "Research challenges and applications for underwater sensor networking," in *2006 IEEE Wireless Communications and Networking Conference, (WCNC '06)*, pp. 228–235, April 2006.

[10] J. Xiong, M. R. Lyu, and K. W. Ng, "Mitigate the bottleneck of underwater acoustic sensor networks via priority scheduling," in *Proceedings of the 6th International Conference on Mobile Ad-hoc and Sensor Networks, (MSN '10)*, pp. 53–60, 2010.

[11] M. Marti, B. Kusy, G. Simon, and G. K. Ldeczi, "The flooding time synchronization protocol," in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems*, pp. 39–49, November 2004.

[12] A. A. Syed and J. Heidemann, "Time synchronization for high latency acoustic networks," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, April 2006.

[13] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," in *Proceedings of the 21st International Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, pp. 1567–1576, 2002.

[14] IEEE STD 802.11-2007, "Wireless LAN medium access control and physical layer specifications," 2007.

[15] J. Deng, B. Liang, and P. K. Varshney, "Tuning the carrier sensing range of IEEE 802.11 MAC," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '04)*, pp. 2987–2991, December 2004.

[16] X. Guo, M. R. Frater, and M. J. Ryan, "A propagation-delay-tolerant collision avoidance protocol for underwater acoustic sensor networks," in *Proceedings of the MTS/IEEE Oceans*, 2006.

[17] J. Ahn and B. Krishnamachari, "Performance of a propagation delay tolerant ALOHA protocol for underwater wireless networks," in *Proceedings of the International Conference on Distributed Computing in Sensor Systems (DCOSS '08)*, 2008.

[18] B. Benson, Y. Li, R. Kastner et al., "Design of a low-cost, underwater acoustic modem for short-range sensor networks," in *Proceedings of the Oceans*, 2010.

[19] L. Freitag, M. Grund, S. Singh, J. Partan, P. Koski, and K. Ball, "The WHOI micro-modem: an acoustic communications and navigation system for multiple platforms," in *Proceedings of the IEEE Oceans Europe*, September 2005.

[20] http://www.isi.edu/nsnam/ns/.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 495127, 11 pages
doi:10.1155/2011/495127

Research Article

A Scalable MAC Protocol Supporting Simple Multimedia Traffic QoS in WSNs

JeongSeok On, HahnEarl Jeon, and Jaiyong Lee

Department of Electrical and Electronics Engineering, Yonsei University, Seoul 120-749, Republic of Korea

Correspondence should be addressed to Jaiyong Lee, jyl@yonsei.ac.kr

Received 15 February 2011; Revised 20 April 2011; Accepted 25 May 2011

Wireless sensor networks (WSNs) are a vibrant research field for the last several years. Especially, design of MAC protocols supporting multimedia traffic QoS is recent challenging work. However, the mechanisms of protocols that can guarantee multimedia traffic QoS have several problems in multihop wireless sensor networks. There are a lot of reasons, including limited end-to-end delay and high throughput. In this paper, we propose ascalable MAC protocol that guarantees multimedia traffic, still images, and scalar sensor data QoS in multi-hop wireless sensor networks. The MAC protocol has made it possible to transmit multimedia streaming traffic without a great change of the end-to-end delay for a specific time. And if there are the some link failures in multi-hop transmission, the protocol can recover the links quickly. Using this algorithm, the proposed MAC protocol outperforms the IEEE 802.11e EDCF and the IEEE 802.15.4 MAC protocol in terms of the end-to-end delay and stable transmission of multimedia streaming data.

1. Introduction

Social concern about applications using multimedia traffic has been growing for the last several years. Research papers concerning applications that generate multimedia traffic, namely, video and audio streaming, still images, and scalar sensor data, are being carried out to implement a communication protocol in the real world. These applications need mostly high throughput and low delay. Networks that connect wireless devices to obtain multimedia traffic are defined as wireless multimedia sensor networks (WMSNs) [1]. In WMSNs, MAC protocols supporting multimedia traffic QoS are divided into three types. The first is contention-based MAC protocols. The typical contention-based MAC protocol is IEEE 802.11e enhanced distributed coordination function (EDCF) that gives the differential priorities according to types of traffic. In EDCF protocol, nodes should participate in contention for channel acquisition through CSMA/CA technique. However, this scheme cannot guarantee multimedia traffic QoS, because the EDCF results in high collision rate at high multimedia traffic load. Most contention-based schemes [2, 3] cannot guarantee multimedia traffic QoS due to collisions of control packets. The second is time division multiple access (TDMA) MAC protocols. TDMA MAC protocols have been studied for supporting multimedia traffic QoS in WSNs. However, the centralized system is used to assign time slots in most TDMA MAC protocols. Due to the centralized system, TDMA MAC protocols [4] are not suitable in WMSNs in terms of the scalability. The third is hybrid MAC protocols. Hybrid MAC protocols [5] using the merit of contention-based and contention-free MAC protocols have been brought to public attention. Due to these features, hybrid MAC protocols can solve the high collisions of packets and scalability problems. Almost all of hybrid MAC protocols are operated in two periods, which are contention period and contention-free period. Hybrid MAC protocols reserve time slots in contention period and nodes transmit data during an assigned slot time in contention-free period. In this protocol, an important factor is the methods that reserve time slots and maintain reservation table in all nodes. If the neighbor nodes do not maintain the same reservation table, collisions of control packets can occur. There have been a few studies that try to support multimedia traffic QoS and maintain reservation table in WMSNs.

2. Related Works

First, we will examine a representative contention-based MAC protocol, namely, IEEE 802.11e EDCF [6, 7]. IEEE 802.11e EDCF uses a priority method to support multimedia

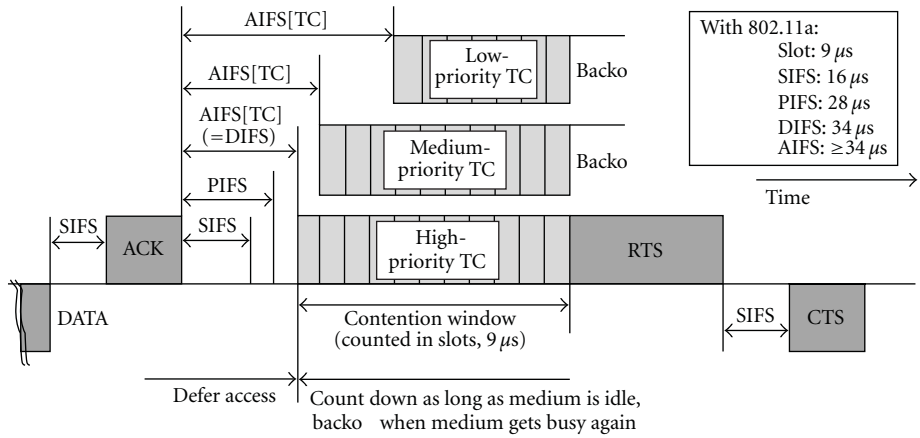


Figure 1: Multiple backoff of EDCF with different priorities.

traffic QoS. The QoS supporting is realized through the introduction of traffic categories (TCs). In Figure 1, differential arbitration inter frame space (AIFS) is applied to every packet according to the access category (AC) of packets. The nodes independently start a backoff after detecting the channel being idle for an AIFS. After waiting for AIFS, each backoff sets a counter to a random number drawn from the interval $[1, CW + 1]$. CWmin refers to the minimum size of the CW. The maximum size of CW is CWmax. EDCF protocol allocates differential values at CWmin and CWmax in accordance with the priority of packets. Each frame from the higher layer arrives at MAC layer along with a specific priority value. Then, each QoS data frame carries its priority value in the MAC frame header. An 802.11e stations shall implement four access categories (ACs), where an AC is an enhanced variant of the DCF 0. Each frame arrives at MAC layer with a priority is mapped into an AC as shown in Table 1. The values of AIFS, CWmin, and CWmax, which are referred to as the EDCF parameters, are announced by the access point (AP) via beacon frames. The AP can adapt these parameters dynamically depending on network conditions Basically, the smaller, AIFS and CWmin, the shorter the channel access delay for the corresponding priority, and hence the more capacity share for a given traffic condition. However, the probability of collisions increases when nodes are operating with smaller CWmin. After all, EDCF is not suitable in multimedia traffic because of high collision rate at high multimedia traffic load.

Second, we will introduce a PARMAC (power-aware reservation-based MAC) [8] that is a major hybrid MAC protocol. This protocol assumes that all stations are synchronized, using a global synchronization scheme, and time is divided in frames of fixed length. Each frame is divided as shown in Figure 2 into the reservation period (RP), during which nodes contend to establish new reservations or cancel reservations, and the contention-free period (CFP), during which they send data packets without contention during the reserved transmission windows and sleep when they do not have packets to send or receive. The lengths of a frame, CFP and RP, are prespecified in the network. Each

Table 1: EDCF priority to Access Category Mappings.

Priority	Access Category (AC)	Designation (Informative)
1	0	Best effort
2	0	Best effort
0	0	Best effort
3	1	Video probe
4	2	Video
5	2	Video
6	3	Voice
7	3	Voice

node operates distributed reservations scheme, namely, each node only attains the link information of 2-hop neighbor nodes. Because of this mode, the scalability of the network can be guaranteed. Basically, PARMAC is appropriate for the transmission of multimedia streaming traffic. However, PARMAC is not suitable for transmission of a periodic scalar sensor data. Besides, correct reservation tables should be maintained by the neighbor nodes. Otherwise, the data that is sent by the reserved nodes might have collisions with several sending data in CFP. Specially, because wireless condition has frequent channel errors, the new reservations do not notice the neighbor nodes exactly due to collision of control packets in the RP period. Finally, even if one node among nodes on the multihop routing path has problem with the data transmission, a source node cannot transmit the data to a destination node. PARMAC does not ensure multimedia streaming traffic QoS when wireless link is broken in certain region during the multimedia traffic transmission.

We also introduce a representative hybrid MAC protocol. The MAC protocol is IEEE 802.15.4 MAC [9] protocol. Most of the pertinent features of IEEE 802.15.4 are in the beacon-enabled mode. The beacon-enabled mode uses a superframe structure where the MAC frame is split in an active and an inactive period; tuning the length of each of them allows to adjust the duty cycle dynamically. The active period consists

of (1) the beacon, (2) a contention access period (CAP), and (3) a collision-free period (CFP). The CFP is only available if guaranteed time slots (GTSs) are allocated by the coordinator; each GTS can occupy multiple timeslots among the 16 available, but only 7 GTSs are allowed in the CFP. A node therefore listens to the beacon first to understand whether a GTS has been reserved by the coordinator or not. If it has, then it remains powered off until its GTS is scheduled to transmit the data. If no GTS is reserved, then it uses CSMA/CA during CAP where the typical backoff procedures are applied.

3. Proposed Schemes

We assume that sensor nodes are high-end devices and all nodes are synchronized at initial stage. High-end devices have the camera that generates multimedia streaming data and still images. All nodes share the location information of the neighbor nodes despite the hello messages. Reservation table is empty initially and routing path is established later. In the following subsection, we will discuss the frame structure of a Scalable QoS (SQ) MAC protocol.

3.1. Frame Structure. The proposed MAC protocol is divided into four periods. As shown in Figure 3, the first is the synchronization period. The second is the random access period. The third is the switching period. Finally, the fourth is the adaptive scheduled access period. These periods will be examined later in detail. We will discuss the way how three traffics are transmitted. We investigate the case that multimedia streaming traffic is generated. In this case, a node operates the reservation set procedure at the random access period and reserves a time slot in the adaptive scheduled access period. All the neighbor nodes of the sender and receiver must record the reservation in their own reservation tables. Once reserved, the sender transmits data and receives ACK message at the adaptive scheduled access period until the reservation cancel procedure is operated. The reservation set procedures occur in nodes along the routing path. As a result, nodes reserve the time slots successively in multi-hop. Nodes operate the reservation cancel procedure when multimedia streaming traffic generation is over. The procedure changes the reserved slots of the nodes' reservation table into free slots at the random access period. Afterward all neighbor nodes can use the time slots for packet transmission.

Next, we will discuss the case of still images and scalar sensor data. These traffics are transmitted in the procedure of RTS/CTS/DATA/ACK at the random access period through the CSMA/CA technique. Moreover the traffic scan be transmitted in the way as slotted CSMA/CA if there are free slots in the adaptive scheduled access period.

3.1.1. Operation in Frame Structure. We will briefly review the operation of each period. First of all, every node is synchronized by the efficient flooding scheme at the synchronization period. In the random access period, nodes must utilize CSMA/CA technique as IEEE 802.11 MAC for channel acquisition. Two kinds of packets are transmitted

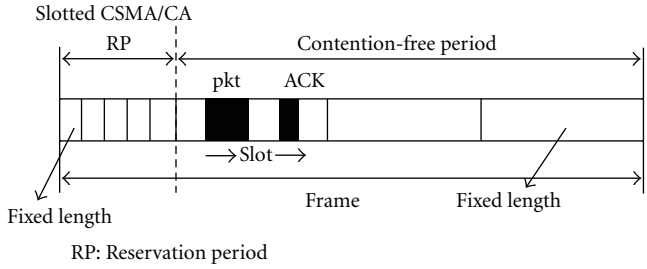


Figure 2: Operation of the PARMAC.

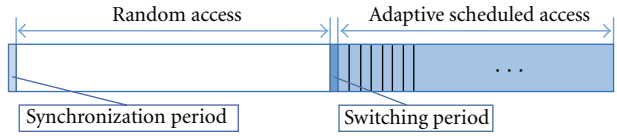
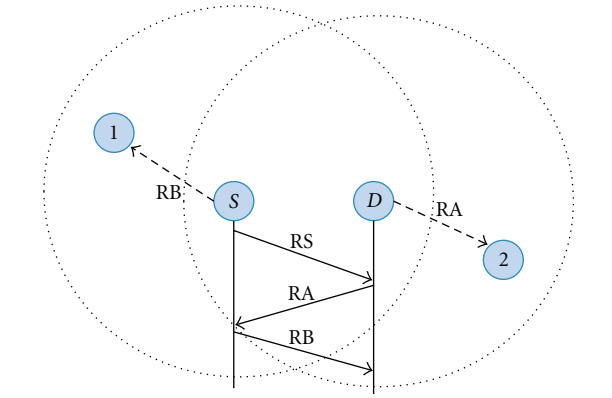


Figure 3: Frame structure of SQ-MAC.

in this period. The first one is reservation control packets. These packets are used for updating reservation table of each node. We will discuss the reservation control packets and the recording method of reservation table in the following chapter. The other is packets for still images and scalar sensor data which are transmitted by the way of RTS/CTS/DATA/ACK. More important packets, such as reservation control packets, are allocated in differential interface space and small contention window size because the packets are transmitted through contention. This will be presented later in detail. In the switching period, the designated nodes must broadcast the slot information of the adaptive scheduled access period. The slot information represents whether the time slot is reserved. If the adaptive scheduled access period is divided into 16 time slots, the information can be represented as 16 bits. In the adaptive scheduled access period, each sending node sends the data and receives ACK message according to the order in the reservation table. If there are free slots in reservation table, all nodes can contend to acquire the slots by slotted CSMA/CA technique. The winner node operates the reservation procedures or sends still images and scalar sensor data. Channel throughput is improved in this way.

3.2. Reservation Procedures. In this section, we will discuss the reservation procedures in detail. The procedures start by the node that generates multimedia streaming traffic. These procedures are operated in the random access period.

3.2.1. Reservation Set Procedure. Reservation procedures can be classified into three operations. The first is the reservation set procedure. This operation starts once when multimedia streaming traffic is generated. Reservation will be maintained until reservation cancel procedure is done. As shown in Figure 4(a), a source node transmits a reservation set (RS) packet to a destination node. The type represents the reservation set packet and the following two fields indicate the addresses of the source and destination node. The slot information field expresses whether each time slot



(a)

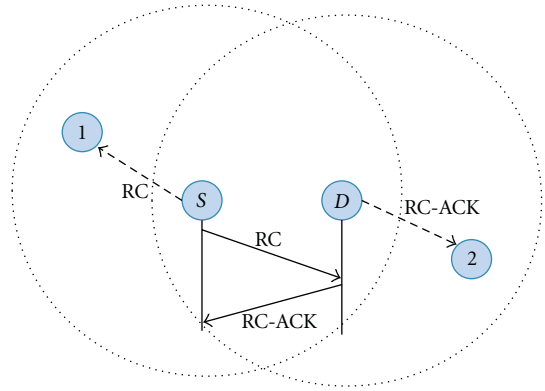
Type	SourceAddr	Dest.Addr.	Slot infor.	Deadline

(b)

Type	SourceAddr	Dest.Addr.	Reserved slot infor.	Time out

(c)

Figure 4: (a) Reservation set procedure. (b) Reservation set format. (c) Reservation ACK and broadcast formats.



(a)

Type	SourceAddr	Dest.Addr.	Reserved slot infor.

(b)

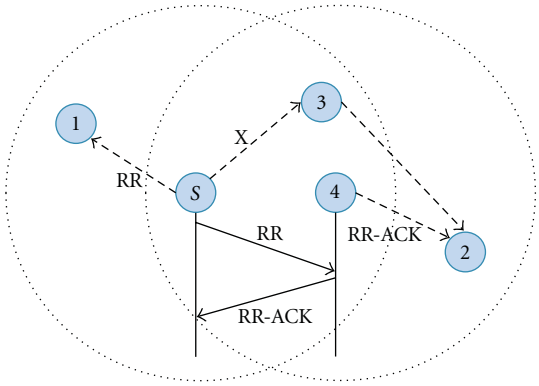
Figure 5: (a) Reservation cancel procedure. (b) Reservation cancel and reservation cancel ACK formats.

of the sender is reserved. If the adaptive scheduled access period consists of 16 time slots, the slot information field is represented as 16 bits and each bit expresses the reservation information of each time slot. The bit 0 indicates a free slot and the bit 1 manifests a reserved slot. Finally, the deadline field represents the remaining time left to reach the destination node. The destination node that receives the reservation set packet compares its own slot information with the reservation table of the source node. If the first 0 match is found in the binary strings in both of these slot information fields, that matching bit slot must be reserved. As an example, if the slot information of the source node is 01110111 and the slot information of the destination node is 11000000, the fifth slot will be reserved. The reservation is corresponded by the destination node with reservation ACK (RA) packet. The reserved slot information field represents 00001000 bits in this case and informs the source node that the fifth slot is reserved. The timeout field indicates the initial value of counter. This field will be examined in detail further in the next subsection. The source node that receives the reservation ACK packet has to broadcast the reservation broadcast (RB) packet. Both the reservation ACK and reservation broadcast packets will inform all neighbor nodes of the related information: the reserved slot, the source address, and the destination address. All nodes have to record the information in their own reservation table.

3.2.2. Reservation Cancel Procedure. The reservation cancel procedure cancels the reservation of the reserved slot in reservation table of the relevant nodes. This procedure will be initiated by the source node that has no data left to be sent or the node that has expired the timeout of the reservation table. Figure 5(a) expresses the procedure. The source node sends the reservation cancel packet and receives the reservation cancel ACK packet from the destination node. Figure 5(b) indicates the format of the packet. All nodes that receive the packets compare the information of the packet with the information of their own reservation table. If the information includes address of the nodes in their reservation table, the nodes will cancel the reservation in their own reservation table.

3.2.3. Reservation Recovery Procedure. It is necessary to resolve the problem when the multimedia streaming traffic QoS cannot be guaranteed, in such cases when at least one node is dead or the channel condition is bad in the special region. The reason is that the transmission is accomplished through multihop transmission. In order to recover this problem quickly, we will need the reservation recovery procedure. Once the reservation set procedure has been completed, the source node sends data and receives the ACK message during the reserved slot time in the adaptive scheduled access period. If the source node does not receive the ACK message during the reserved slot time, the source node will recognize connection failure. Then the source node will find the other destination node. As a general rule, the selected node will be located close to the original node. Then the reservation recovery procedure will proceed. The source node has to transmit the reservation recovery packet to the selected node and receive the reservation recovery ACK packet. The packet fields indicate the address of the original node in the lost address field and address of the selected node in the destination address field. All nodes that overhear the packets have to change the lost address to destination address in reservation table.

Owing to this procedure, the problem can be recovered rapidly. For example, as shown in Figure 6(a), the multimedia traffic flow is $1 \rightarrow S \rightarrow 3 \rightarrow 2$ and the situation



(a)

Type	SourceAddr	Dest.Addr.	Lost Addr.	Reserved slot infor.	Time out

(b)

Slot number	Source node	Dest. node	Timeout (ms)	Slot infor.
1	S	3	12	1
2	3	2	10	1

(c)

Slot number	Source node	Dest. node	Timeout (ms)	Slot infor.
1	S	4	20	1
2	4	2	20	1

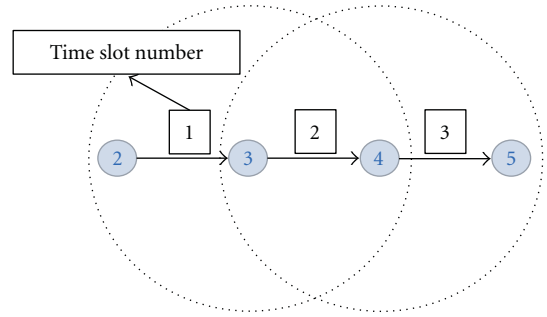
(d)

Figure 6: (a) Reservation recovery procedure. (b) Reservation recovery and reservation recovery ACK formats. (c) Reservation table of all nodes before the reservation recovery procedure. (d) Reservation table of all nodes after the reservation recovery procedure.

is that the reservations have been completed. When the connection between the nodes S and 3 breaks down, the node S must operate the reservation recovery procedure. The node S selects a node 4 as a forwarding node instead of the node 3. The connection can be recovered rapidly. Figures 6(c) and 6(d) show the reservation table of nodes before and after the reservation recovery procedure.

3.3. Reservation Table. The reservation table can be updated by listening and overhearing the reservation control packets in the random access period. All the nodes that listen to the reservation ACK or reservation broadcast have to register the address of the source and destination node and the reserved slot number in their reservation tables. The timeout counter is also initiated by the predetermined value. Consider Figures 7(a) and 7(b) for example. The routing path is 2 → 3 → 4 → 5. The reservations have been completed during the random access period through the reservation set procedure. The value between the nodes indicates the number of the reserved slot. Figure 7(b) shows the reservation table of the node 3. The slot information consists of 0 and 1 bits, and the bit 1 represents a reserved slot and the bit 0 indicates a free slot. Figure 8 describes the slot information in detail.

The slot information is very important for the reservation among the nodes and escaping collision. This slot



(a)

Slot number	Source node	Dest. node	Timeout (ms)	Slot infor.
1	2	3	10	1
2	3	4	15	1
3	4	5	20	1
4				0

(b)

Figure 7: (a) Network topology. (b) Reservation table of node 3.

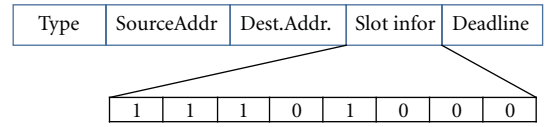


Figure 8: Slot information field.

information is utilized in the switching period. In the next subsection, we will describe the switching period.

3.4. Contention Method of the Random Access Period. In the random access period, reservation control packets and the RTS packets for sending still images and scalar sensor data will contend for channel acquisition. We will give priority to the reservation control packets than the RTS packet. The method that gives priority to the reservation control packets sets up the differential interface space (IFS) and contention window size. Table 2 shows the parameters of interface space and contention window size of the control packets.

3.5. Description of the Switching Period. Due to unstable wireless channel and collision among the reservation control packets, the reservation table of a node and the neighbor nodes cannot coincide. In this case the multimedia traffic transmission can generate collision with each other. As a result, multimedia traffic QoS cannot be guaranteed. In order to solve this problem, the specific nodes broadcast the slot information during the switching period. The specific nodes that send or receive the reservation control packets during the prior random access period should broadcast the slot information according to the random backoff scheme. The exact reservation table will be maintained among the neighbor nodes (Figure 9).

3.6. Description of the Adaptive Scheduled Access Period. The adaptive access period consists of several fixed time slots. The

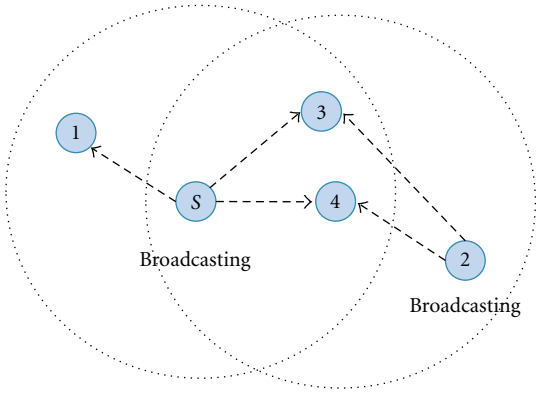
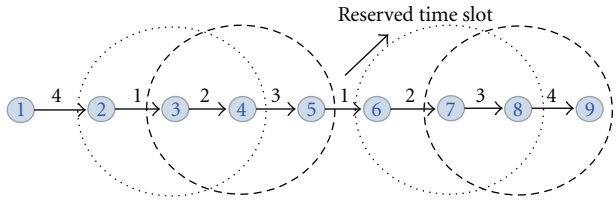


Figure 9: Broadcasting slot information.



(a)

Slot number	Source node	Dest. node	Timeout (ms)	Slot infor.
1	2	3	10	1
2	3	4	15	1
3	4	5	20	1
4	1	2	10	1

(b)

Slot number	Source node	Dest. node	Timeout (ms)	Slot infor.
1				0
2	6	7	10	1
3	7	8	8	1
4	8	9	15	1

(c)

Figure 10: (a) Network topology for simulation. (b) Reservation table of node 3. (c) Reservation table of node 8.

length of a time slot is determined by the network manager. If data is transmitted successfully during each time slot, the nodes will reset the timeout value in the reservation table.

In case when the timeout value is 0, the node shall perform the reservation cancel procedure. The number of time slots should be limited according to the volume of the traffic. During the free time slots, all forwarding nodes can transmit data through a slotted CSMA/CA technique. Nodes without sending or receiving data can sleep during the time slots. Figure 10 shows an example of multihop transmission for multimedia traffic.

4. Performance Analysis

We will examine the performance analysis under the N-hop linear topology. Figure 11 shows the topology. We will find both the optimal random access period and

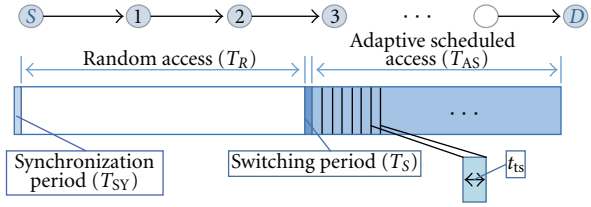


Figure 11: Network topology and analysis variables.

adaptive scheduled access period for minimum end-to-end delay according to the multimedia streaming data rate. In Figure 11, node S is a source node that generates multimedia streaming traffic. Node S delivers multimedia streaming traffic to a destination node through multihop transmission for a while. The objective function refers to the expected time of the end-to-end delay from node S to node D. Variables of the optimization problem indicate the number of time slots in the adaptive scheduled access period and the length of the random access period:

$$\begin{aligned} \text{minimize} \quad & E\left[\text{Delay}_{\text{e2e}}\right] \\ \text{subject to} \quad & T_R \geq 0, \\ & n \geq 0, \ \text{integer}. \end{aligned} \tag{1}$$

We assume no processing delay of CPU, no queueing delay, and no propagation delay. The synchronization period and switching period are constant values.

Table 3 shows the analysis parameters for the performance evaluation.

We defined a frame time as the sum of the four periods, which are the synchronization period, random access period, switching period, and adaptive scheduled access period. The following equation expresses a frame time:

$$T_{\text{frame}} = T_{\text{SY}} + T_R + T_S + T_{\text{AS}}. \tag{2}$$

The adaptive scheduled access period is divided into n time slots. Therefore, the adaptive scheduled access period is expressed as the product of n and a time slot. The following equation shows the numerical formula:

$$T_{\text{AS}} = n \cdot t_{\text{ts}}. \tag{3}$$

One time slot should be enough time to receive and send a data packet. As shown in (4), a time slot indicates the sum of the data transmission time, ACK receiving time, three times of short interframe space time, and guard time:

$$t_{\text{ts}} = \frac{L_{\text{data}}}{t_B} + t_{\text{ACK}} + 3 \cdot t_{\text{SIFS}} + \text{guard\_time}. \tag{4}$$

Now, we will examine the expected time of a reservation set procedure. This procedure proceeds in the random access period. In other words, the CSMA/CA technique is used for channel acquisition. First, we wait for the IFS time after the preceding transmission. The IFS time for the reservation set procedure appears at Table 2.Then, the random backoff

Table 2: Parameters of interface space and contention window.

Control packet	IFS	CWmin	CWmax
Reservation set	2*Slot_time+SIFS	15	31
Reservation cancel	2*Slot_time+SIFS	15	31
Reservation recovery	2*Slot_time+SIFS	7	15
RTS (still images and data)	3*Slot_time+SIFS	31	127

Table 3: Analysis parameters.

Random access period	T_{R}
Synchronization period	T_{SY}
Adaptive scheduled access period	T_{AS}
Switching period	T_{S}
The hop number	N_{h}
The number of the time slots	n
The multimedia streaming data rate	r_{data}
The data length	L_{data}
Time to Tx/Rx a byte	t_{B}
The time of a time slot	t_{ts}
The time of a ACK packet	t_{ACK}
The time of the contention window	t_{CW}
The time of the RS	t_{RS}
The time of the RA	t_{RA}
The time of the RB	t_{RB}
The time of the IFS	t_{IFS}
The time of the SIFS	t_{SIFS}
The time of RS procedure	t_{RSP}

time for contention follows the IFS time. The average value of the random backoff time is half of contention window size. After sending the reservation set, reservation ACK and reservation broadcast packets are transmitted by the source and destination nodes. Two short interframe spaces are among these packets. The following equation indicates the expected time of a reservation set procedure:

$$E[t_{\mathrm{RSP}}] = t_{\mathrm{IFS}} + \frac{t_{\mathrm{CW}}}{2} + t_{\mathrm{RS}} + t_{\mathrm{RA}} + t_{\mathrm{RB}} + 2 \cdot t_{\mathrm{SIFS}}. \quad (5)$$

The maximum number of the completed reservation set procedure in a frame period, $\mathrm{MN}_{\mathrm{RSP}}$, is expressed in (6). Reservation set procedures must be performed in the random access period. So the maximum number of the completed reservation set procedures is the value of random access period divided by the expected time of a reservation set procedure. Because this value is integer, the decimal fraction is erased:

$$\mathrm{MN}_{\mathrm{RSP}} = \left[\frac{T_R}{E[t_{\mathrm{RSP}}]}\right]. \quad (6)$$

Now, we find the probability of one more hops' transmission during a frame period. It is possible that the sending node reserves a preceding time slot than a time slot of the relay node in the adaptive scheduled access period. The probability, P_r, is expressed in (7). We assume that the number of the unreserved time slots is k. If the sending node selects a time slot, the relay node chooses one of the following time slots than the selected time slot. The probability is independent of the value of k. The value of k is an invariable in

$$P_r = \frac{1}{k}\left(\frac{k-1}{k-1} + \frac{k-2}{k-1} + \cdots + \frac{2}{k-1} + \frac{1}{k-1}\right) = \frac{1}{2}, \quad 0 \le k \le n. \quad (7)$$

We will use the probability of one more hops' transmission to find the average hop count of the relayed packets during a frame period. Because the probability of one more hops' transmission is the same regardless of the value of k, the average hop count of the relayed packets is expressed as the sum of the probability from one hop transmission to the maximum hops transmission during a frame period:

$$E[M_h] = 1 \cdot 1 + 2 \cdot P_r + \cdots + \mathrm{MN}_{\mathrm{RSP}} \cdot (P_r)^{\mathrm{MN}_{\mathrm{RSP}}-1}. \quad (8)$$

We now examine the average end-to-end delay in the multihop transmission condition. First, we find the average number of a frame period spent for multihop transmission. The value is that the hop number from a source node to a destination node is divided by the average hop count of the relayed packets. The average end-to-end delay equals the product of the average number of a frame period and a frame time. The following equation shows the average end-to-end delay:

$$E\left[\mathrm{Delay}_{\mathrm{e2e}}\right] = \left[\frac{N_h}{E[M_h]} + 1\right] \cdot T_{\mathrm{frame}}. \quad (9)$$

We assume no queueing delay. For this we apply $D/D/1$ queueing model in sensor nodes. Both the arrival and the service rate are constant value. The arrival rate is smaller than the service rate. However, there is waiting delay due to channel contention. If a sensor node transmits data, the two forwarding nodes cannot send the data in Figure 10.The constraint function of the random access period is indicated in (10). The minimum random access period is the product of the expected time of a reservation set procedure and the three times the number of the generated data during a frame time:

$$T_R \ge 3E[t_{\mathrm{RSP}}] \cdot (r_{\mathrm{data}} \cdot T_{\mathrm{frame}}). \quad (10)$$

The number of the time slots is expressed as (10). The minimum number of the time slots is three times the number of the generated data during a frame time. The number is integer:

$$n \ge [3r_{\mathrm{data}} \cdot T_{\mathrm{frame}} + 1]. \quad (11)$$

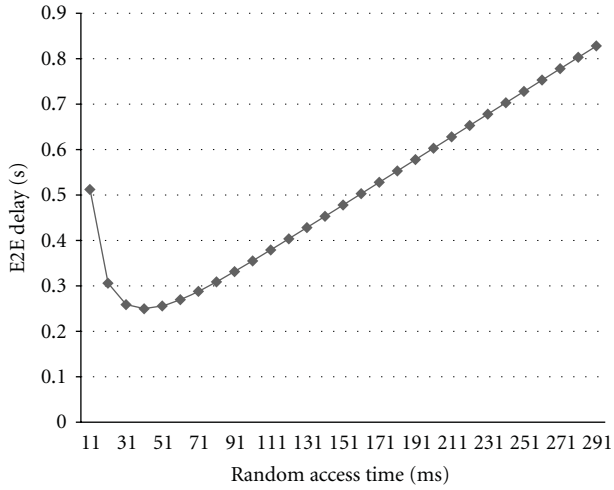


Figure 12: E2E delay when the data rate is one.

Now, we define the end-to-end delay as an optimization problem through the previous equations. The following equation indicates the optimization problem:

$$\begin{aligned} \text{minimize} \quad & E\left[\text{Delay}_{\text{e2e}}\right] \\ \text{subject to} \quad & T_R \geq 3E[t_{\text{RSP}}] \cdot (r_{\text{data}} \cdot T_{\text{frame}}), \\ & n \geq [3r_{\text{data}} \cdot T_{\text{frame}} + 1], \\ & T_R \geq 0, \\ & n \geq 0, \ \text{integer}. \end{aligned} \tag{12}$$

We will find the optimal random access period and the number of time slots in each senor node according to the data rate from the source node. Therefore, we arrange (12) for the math problem in terms of a random access period and the number of time slots. The following equation shows the final optimization problem of the end-to-end delay:

$$\begin{aligned} \text{minimize} \quad & E\left[\text{Delay}_{\text{e2e}}\right] = \left[\frac{N_h}{E[M_h]} + 1\right] \\ & \cdot (T_{\text{SY}} + T_R + T_S + n \cdot t_{\text{ts}}) \\ \text{subject to} \quad & \left(\frac{1}{3E[t_{\text{RSP}}] \cdot r_{\text{data}}} - 1\right) \cdot T_R - t_{\text{ts}} \cdot n \geq T_{\text{SY}} + T_S, \\ & \left(\frac{1}{3r_{\text{data}}} - t_{\text{ts}}\right) \cdot n - T_R \geq T_{\text{SY}} + T_S, \\ & T_R \geq 0, \\ & n \geq 0, \ \text{integer}. \end{aligned} \tag{13}$$

We solved (13) using mathematical program and found the optimal values as the data rate varies. The parameters except three variables use the values of Table 4. The distance from a source node to a destination node is ten hops apart. When data rate is one unit per second, the graph of end-to-end delay is convex function as the random access period varies as shown in Figure 12. Even though the data rate varies, the graph of the end-to-end delay is almost the same pattern in terms of the random access period. Therefore the random access period that satisfies the constraints and minimum end-to-end delay is a global optimization value.

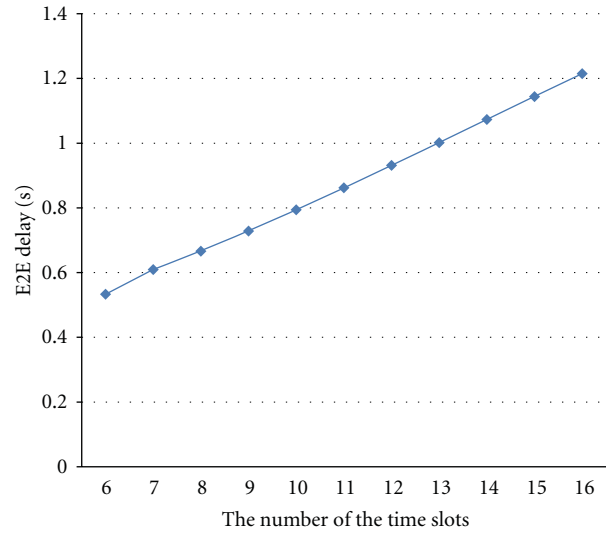


Figure 13: E2E delay as the number of the time slots varies when the data rate is ten.

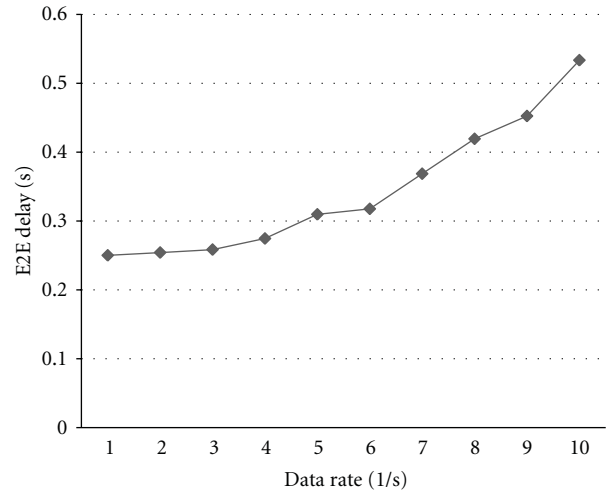


Figure 14: Optimal E2E delay as the data rate varies.

Figure 13 shows the end-to-end delay as the number of the time slots varies when the data rate is ten. More the number of the time slots than optimal number of those result in increasement of the end-to-end delay. Therefore the optimal number of the time slots in adaptive scheduled access period is important for multimedia traffic QoS.

As shown in Figure 14, the end-to-end delay of multimedia streaming traffic increases linearly as the data rate rises. The reasons of the linear delay increasement are that both the number of time slots in the adaptive scheduled access period and the random access period have to increase according to the data rate. Once the reservation set procedure

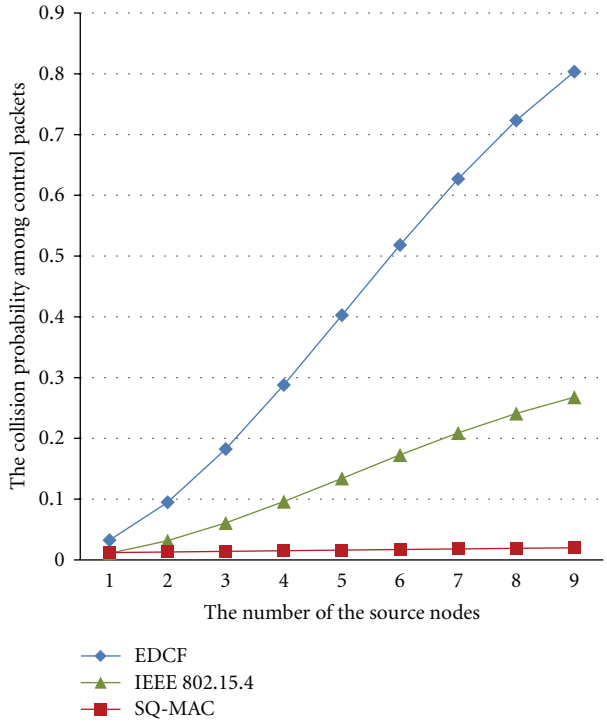


Figure 15: The collision probability among control packets as the number of the source nodes increases from one to nine.

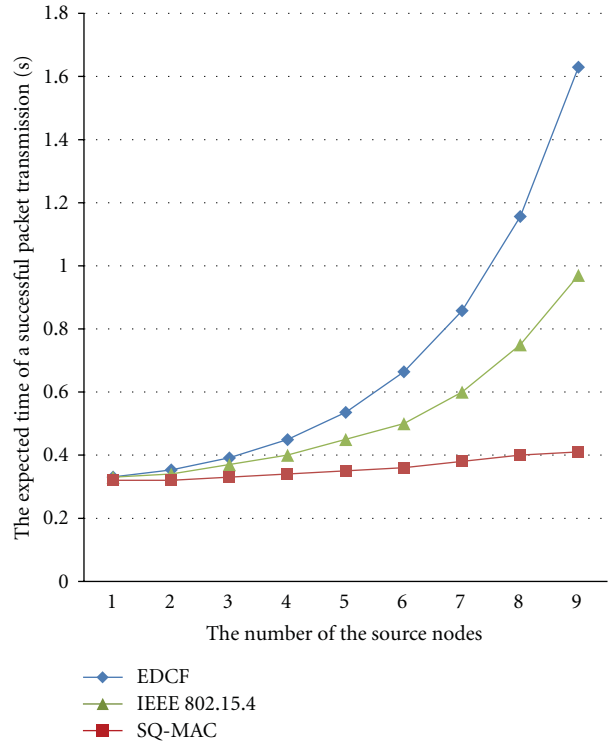


Figure 16: The expected time of a successful packet transmission as the number of the source nodes increases from one to nine.

has completed in nodes on multihop path from a source node to a destination node, the reserved slots will be used to guarantee the multimedia streaming traffic QoS until the reservation cancel procedures have been operated by the node. As a result, the number of the control packets for several reservations in each node is much less than those of other protocols. The collisions of the control packets have hardly occured in SQ-MAC as we expected.

5. Simulation Results

The network topology for the simulation is represented in Figure 11. The simulation parameters are summarized in Tables 2 and 4. We used the MaTLAB simulator for the simulation. The distance from node S to node D is 10 hops away in the network topology. We assume that both the synchronization period and the switching period are one millisecond, respectively. The beacon period is also one millisecond in IEEE 802.15.4 MAC protocol. The environment condition is that multimedia streaming traffic and still images are generated at the constant packet rate (CPR). The number of the source nodes on routing path increases from one to nine. In this case we examined the expected packet delay from node S to node D until the transmission of one thousand packets from node S has been completed. We will compare the SQ-MAC with IEEE 802.11e EDCF and IEEE 802.15.4 MAC protocols in terms of the several variables. IEEE 802.15.4 MAC protocol operates in beacon-enabled mode. The coordinator nodes are 1, 4, 7 and destination nodes. Throughout the nodes, guaranteed time slots is reserved for multimedia streaming traffic. In beacon period, the coordinator nodes notice the reserved time slots.

First, we observed the collision probability of the control packets for data transmission in three protocols as the source node increases from one to nine nodes. The control packets include RTS, reservation set, and polling packets. The collision probability of the control packets increases rapidly because of the exponential growth of control packets for transmission of burst multimedia traffic in EDCF protocol (see Figure 15). The probability almost approaches 80 percent in the case of nine source nodes. Thus, most of the RTS packets must be retransmitted by source nodes in order to utilize wireless channel. On the other hand, the collision probability of the control packets increases slowly in the IEEE 802.15.4 MAC and SQ-MAC protocols. The reason is that the protocols hardly use control packets for multimedia traffic transmission.

After the successful control packet transmission, the data packet is transmitted by the node. Therefore the expected transmission time of a data packet has relevance to the retransmission number of the control packets. We define the reservation set packets as the control packets in SQ-MAC. There is almost no increase in the number of the control packets as multimedia streaming traffic increases in SQ-MAC. The reason is that only one control packet is transmitted to reserve a time slot when the node starts to transmit multimedia streaming traffic. Therefore, the collision probability of the reservation set packets is much lower than that of the RTS packets in EDCF protocol. In EDCF protocol, the expected time of a successful packet

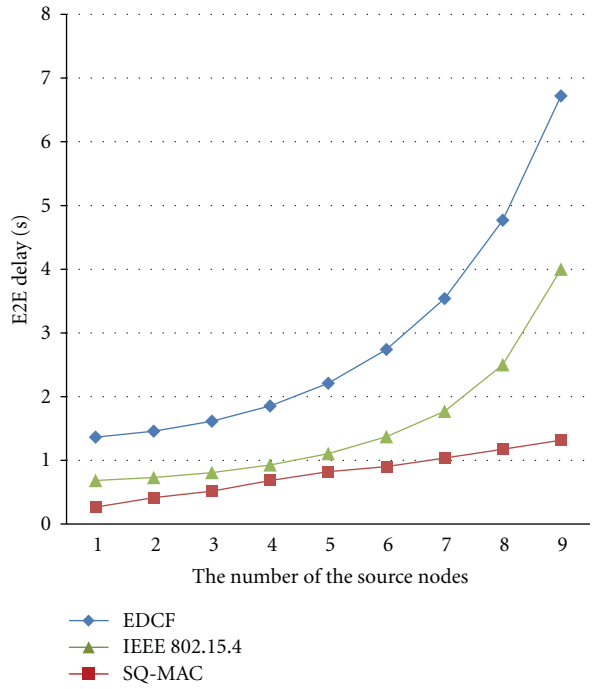


Figure 17: E2E delay as the number of the source nodes increases from one to nine.

transmission has increased rapidly as the number of the source nodes increases due to the exponential collision growth of the RTS packets as shown in Figure 15. As a result, the channel utilization is very low in EDCF protocol. In IEEE 802.15.4 MAC, the control packets are polling packets for GTS reservation and RTS packets using slotted CSMA/CA in CAP. The coordinator nodes cannot use the GTS reservation packets unlike the other nodes. Therefore, IEEE 802.15.4 protocol shows better performance than EDCF protocol (Figure 16).

Finally, we will examine the end-to-end delay from node S to node D as the number of the source nodes increases. In EDCF protocol, the delay increases geometrically because of the retransmission of the RTS packets. For the retransmission of the RTS packets, the node waits for the random backoff time in order to avoid collision. As the number of the retransmission increases, the backoff time grows, which results in the increasement of the end-to-end delay in network topology. Besides, there are wide variations in the end-to-end delay of packets. In other words, EDCF protocol cannot guarantee a stable transmission supporting multimedia traffic QoS. In IEEE 802.15.4 MAC protocol, the delay increases geometrically because of the collision among the RTS and GTS reservation packets. On the other hand, SQ-MAC not only makes shorter end-to-end delay due to the low collision rate of the control packets but also stable multimedia traffic transmission through the reservation of time slots (Figure 17).

6. Conclusion

We proposed SQ-MAC supporting several kinds of multimedia traffic. Owing to the protocol operation in a distributed fashion, SQ-MAC is scalable in WMSNs. In this condition, multimedia traffic packets are transmitted through multihop path. Therefore, we should resolve several problems in order to support multimedia traffic QoS. Even if one node on the routing path is not capable of transmitting multimedia packets, the end-to-end delay increases rapidly. To solve this problem, we propose the reservation recovery procedure. The second problem is that EDCF and IEEE 802.15.4 MAC protocols cannot guarantee multimedia traffic QoS due to the collisions among the control packets. To tackle this problem, we suggest the reservation set procedure that reduces the number of the control packets. It is possible that data packets are transmitted from a source node to a destination at constant intervals by allocating the time slots to multimedia traffic packets. Due to unstable wireless channel or collision among the reservation control packets, the reservation table of a node and the neighbor nodes cannot coincide. To resolve this matter, the designated nodes broadcast the slot information during the switching period.

SQ-MAC protocol contributes to the performance improvement in terms of the end-to-end delay, jitter, and throughput in WMSNs. In the future, we will analyze energy consumption in the proposed scheme and more simulations will be operated.

Table 4: Simulation parameters.

Simulation parameter	
The number of source nodes	k
Channel throughput	3 Mbytes/s
Packet Length	
All control packets	10 bytes
DATA	2.5 kbytes
Traffic Classification	
Multimedia streaming data	CPR 10 units/s
Still image and data	CPR 1 units/s
Interface Space and Contention Window Size	
DIFS	0.022 ms
SIFS	0.016 ms
A contention window slots	0.009 ms
A time slot	2 ms
IEEE 802.15.4 MAC	
Frame period	20 ms
Contention access period	6 ms
Contention free period	14 ms
The number of guaranteed time slots	7
SQ-MAC	
Frame period	20 ms
Random access period	4 ms
Adaptive scheduled period	16 ms
The number of adaptive scheduled slots	8

Acknowledgment

This research was supported by the MKE (The Ministry of Knowledge Economy), Republic of Korea, under the ITRC

(Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-C1090-1111-0006).

References

[1] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Computer Networks*, vol. 51, no. 4, pp. 921–960, 2007.

[2] J. On, J. Kim, J. Lee, Y. Kim, and H. Chong, "A MAC protocol with adaptive preloads considering low duty-cycle in WSNs," in *the 3rd International Conference on Mobile Ad-hoc and Sensor Networks (MSN '07)*, vol. 4864 of *Lecture Notes in Computer Science*, pp. 269–280, Beijing, China, December 2007.

[3] J. Kim, J. On, S. Kim, and J. Lee, "Performance evaluation of synchronous and asynchronous MAC protocols for wireless sensor networks," in *the 2nd International Conference on Sensor Technologies and Applications (SENSORCOMM '08)*, pp. 500–506, Cap Esterel, France, August 2008.

[4] G. Jolly and M. Younis, "Energy efficient collision-free medium access control for wireless sensor networks," *Wireless Communications and Mobile Computing*, 2005.

[5] I. Rhee, A. Warrier, M. Aia, J. Min, and M. L. Sichitiu, "Z-MAC: a hybrid MAC for wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 511–524, 2008.

[6] S. Choi, J. Del Prado, S. Shankar N, and S. Mangold, "IEEE 802.11e Contention-based channel access (EDCF) performance evaluation," in *IEEE International Conference on Communications*, vol. 2, pp. 1151–1156, Anchorage, Alaska, USA, May 2003.

[7] L. Romdhani, Q. Ni, and T. Turletti, "Adaptive EDCF: enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '03)*, pp. 1373–1378, New Orleans, La, USA, March 2003.

[8] M. Adamou, I. Lee, and I. Shin, "An energy efficient real-time medium access control protocol for wireless ad-hoc networks," in *the 8th IEEE Real-Time and Embedded Technology and Applications Symposium*, September 2002.

[9] "IEEE 802.15.4. Wireless Medium Access Control (MAC) and Physical Layer (PHY)Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)," 2003.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 107062, 17 pages
doi:10.1155/2011/107062

Research Article

Quadratic Programming for TDMA Scheduling in Wireless Sensor Networks

Gergely Treplán,[1] Kálmán Tornai,[1] and János Levendovszky[2]

[1] *Faculty of Information Technology, Pazmany Peter Catholic University, P.O. Box 278, 1444 Budapest, Hungary*
[2] *Department of Telecommunications, Budapest University of Technology and Economics, P.O. Box 91, 1521 Budapest, Hungary*

Correspondence should be addressed to Gergely Treplán, trege@itk.ppke.hu

Received 16 February 2011; Revised 22 June 2011; Accepted 29 June 2011

This paper presents a novel Multihop Aperiodic Scheduling (MAS) algorithm which guarantees energy-efficient data collection by Wireless Sensor Networks (WSNs) under delay constraints. Present Medium Access Control (MAC) protocols in WSNs typically sacrifice packet latency and/or the reliability of packet transfer to achieve energy-efficiency. Thus, the paper is concerned with developing a novel protocol to achieve energy efficient and reliable multihop data transfer in WSNs satisfying given latency requirements. Energy efficiency is achieved by optimizing the scheduling of the underlying Time Division Multiple Access (TDMA) system by minimizing the wake-up number of the nodes. Schedule optimization is transformed into a quadratic programming (QP) task, which is then solved by the Hopfield net in polynomial time. In this way, an energy efficient scheduling can be obtained which meets a given delay requirement in TDMA systems. The performance of the new algorithm has been evaluated by simulations and compared to the performance of well-known scheduling methods, such as SMAC, UxDMA (a slot assignment algorithm for WSN), and traditional tree-based protocols. The simulations have demonstrated that our method reduces global power consumption for time-driven monitoring.

1. Introduction

Due to the recent developments in communication technologies and microelectronics, Wireless Sensor Networks (WSNs) are capable of conveying high-resolution information processes to a Base Station (BS) [1, 2]. However, the longevity of such networks has become of crucial importance in applications like environment and health monitoring, where WSNs might have to be operational for several years [3, 4]. Therefore one of the most important problems of WSNs stems from the limited energy storage capacity which imposes severe limits on the longevity [5, 6].

The major energy consumption of a sensor node results from the active state of its radio component [7]. Thus, by "sleeping" (i.e., switching off the radio module), one may save a considerable amount of energy [8]. Therefore the lifespan of a wireless network is roughly defined by how often the electricity consuming parts are turned on. In WSN, the physical layer was designed with short radio range to meet the requirement of low energy consumption. As a result, the network frequently relies on multihop communication schemes to the BS but this procedure leads to asymmetry in the energy consumption; nodes close to the BS are forced to be engaged in higher forwarding activity [6]. Existing routing approaches of load and energy balancing employ periodical route changes [9, 10], however, most of the work does not take into account the underlying channel access method. Hence, the proper design of medium access (MAC) gives rise to new challenges and it is intensively researched [8, 11–13]. A large number of research work have proposed low power MAC protocols for multihop WSNs [14–17], which is going to be detailed in Section 2.

In WSN, the reduction of energy consumption must be one of the primary goals in the design, however, collision avoidance capability can also be a crucial metric. To achieve energy efficiency, we need to identify what are the main sources that cause inefficient use of energy as well as what tradeoffs we can make to reduce energy consumption. The following factors have been identified as the most significant ones regarding energy consumption [18].

Collision. When a transmitted packet is not delivered it has to be discarded and the retransmissions increase energy consumption as well as the latency.

Overhearing Overhead. This is due to the fact that a node accidentally picks up packets that are destined to other nodes.

Idle Listening. When the node listens to receive possible packets, which are not sent.

Control Packet Overhead. Sending and receiving control packets.

Plenty of channel access methods are proposed to overcome these drawbacks. There are two common solutions: (i) contention based; (ii) TDMA solutions. In case of contention-based MACs, nodes try to assign the channel themselves competitively. Since nodes do not cooperate, collisions happen and increase energy consumption and latency severely. However, collision can be eliminated by using four handshaking RTS/CTS strategies, the energy overhead is huge due to the extra energy consumption of control packets. Overhearing and idle listening are also typical in contention-based MACs and consume about half of the energy (one can see more details of exact measurements presented in [19]). These distributed protocols are reactive to environment changes, however they do not exploit the topology, routing, and traffic information, which with more efficient channel assignment could be achieved. In case of stochastic traffic pattern, contention-based solutions transmit the data in a very fast manner to the BS, but it still fails to provide energy efficient communication. On the other hand, the fully cooperative TDMA-based solutions communicate on different time slots to prevent conflicts and offer several advantages for data collection as compared to contention-based protocols. They eliminate collisions, overhearing, and idle listening and the network does not suffer from extra energy overhead (i.e., RTS/CTS overhead). Note, that idle listening inherently appears in TDMA systems, since the time slot has to be greater than the packet duration. On the other hand, they also permit nodes to enter into sleep modes during inactive periods, thus, achieving low duty cycles and conserving energy. From the point of quality of service (QoS) TDMA-based communications can provide provable guarantee on the completion time of data collection, for instance, in timely detection of events. On the other hand, TDMA needs tight synchronization and accurate traffic and topology information, which implies extra communication overhead compared to a contention-based MAC.

In this work, we present a novel scheduling algorithm to unite the advantages of TDMA and contention-based protocols, that is, energy-efficient communication with acceptable latency. Namely, we are seeking an energy optimal schedule for a given traffic demand and routing topology, while a given end-to-end latency requirement is satisfied (i.e., the length of the schedule does not exceed a predefined threshold). This gives a cross layer perspective to our work, as the application and the routing information are used to determine the schedule. This cross layer optimization will be carried out by Binary Quadratic Programming (BQP). BQP is a well-known NP-hard problem, but we solve it by the Hopfield neural network (HNN). However, other solvers (e.g., semidefinite relaxation [20]) can also be used.

The rest of this paper is organized as follows: (i) Section 2 discusses the related works; (ii) Section 3 describes the system model; (iii) Section 4 introduces the optimization problem to be solved; (iv) Section 5 presents the HNN to minimize the penalty functions mapped into quadratic programming and realize feasible schedules; (v) Section 6 evaluates the performance by extensive simulations; (vi) some concluding remarks can be found in Section 7.

2. Related Works

In the recent years, a large number of the energy-efficient MAC schemes in the data link layer have been proposed for WSNs in recent years [8, 21]. Current MAC design for WSNs can be broadly divided into two common solutions, contention based and scheduled based (or TDMA protocols).

In the contention-based scheme, nodes try to access the channel randomly, independently from each other, such as ALOHA [22]. ALOHA has been improved in many ways to achieve different performance requirements (e.g., no collision) by using such methods as the four-way handshaking carrier sense multiple access with collision avoidance (CSMA/CA) [23]. In WSN, B-MAC [24] is the most popular example of the contention-based protocol, and it is mainly built on the well-known standardized protocol for wireless ad hoc networks the IEEE 802.11 [25]. The B-MAC protocol reduces idle listening and provides an asynchronous channel access method, however X-MAC [26] and C-MAC [27] made an improvement on it by trying to use shorter preambles and avoid the overhearing among neighboring nodes. Contention-based protocols are very important channel access methods in the case of rare, but bursty traffic, which is the typical in event driven mode.

TDMA-based packet transmission scheduling schemes could be very efficient if the users are known and their number is fix, the data arrives regularly (e.g., every one minutes the temperature value must be delivered to the BS and topology are static). Time-slotted communication is robust during heavy traffic loads while contention-based access protocols may fail to allocate the medium successfully when the data rates and the number of sources are high. The TDMA-based protocols can also save more energy, since each node can stay in sleep mode and wakes up during its own slot time. For these reasons, TDMA has been a subject of extensive research and there are also several TDMA commercial standards and applications [28]. Scheduled or TDMA-based protocols use topology information as a basis of scheduling. Based on tight synchronicity among their neighbors, time slots are scheduled for access in such a way that no two interfering nodes access the channel at the same time. Note that the synchronization method is out of the scope of this paper, however, one must note that clock drifts can cause energy overhead in TDMA systems (i.e., since synchronicity could only be guaranteed only with a certain accuracy, hence receiver must wake up a certain guard

time before the sender does). The most popular TDMA-based MAC protocols proposed for WSNs are as follows: S-MAC [29] is one of the standard solution in the TinyOS [30] operating system designed to be used with networked sensors and it supports the Mica2 platform [31]. The S-MAC is a method using loose synchronization and RTS/CTS handshaking which occurs significant energy overhead in case of small packets. Furthermore, the paper [32] proposes SMACS as an energy-efficient interference free TDMA-based protocol. A link is only active (i.e., it consumes energy) in its own slot when a packet is generated to be sent and one frame contains all the slots. Hence time slot assignment is the main component of TDMA protocols in wireless networks. Slot assignment problem is NP hard [28]. In paper [33], the protocols RAND, MNF, and PMNF are proposed which provide heuristic solutions to the slot assignment problem. In [34] authors presents DRAND a distributed version of RAND, which runs fast and efficiently. This slot assignment solution uses constant frame size such as SMACS or SMAC, however, a hybrid protocol Z-MAC [35] protocol contains DRAND and adapts the size of the frame. Static frame size is not optimal because of the asymmetry of the data collection tree and traffic pattern [28]. Figure 1 summarizes the typical frame constructions.

In [36] scheduling for QoS routing is presented, where QoS means that traffic requires a given number of slots per frame in the QoS route and formulates it as a slot shortage problem. This paper proposed a distributive solution for the slot shortage problem, which even works in mobile multihop wireless networks. However the protocol presented in this paper assumes static topology and the primary QoS metric is latency instead of data rate. Goldsmith et al. [37] also uses cross-layer-based optimization perspective as Jurdak et al. [38, 39], however, none of them build energy optimal schedules with respect to a given data collection time.

In the rest of the section we analyze the solutions assuming that the topology and the data collection tree are given. A very good summary of such scheduling algorithms for tree networks can be found in [28]. According to [28], schedules can be optimal in terms of the following performance metrics: (1) energy consumption: the energy consumed by the data gathering operation; (2) schedule length/data collection time: the time duration in which the last generated packet is gathered; (3) latency: the maximum end-to-end latency of data gathering; (4) fairness: a metric to determine whether users or applications are receiving a fair share of system resources. Ukai et al. presented a novel formulation for scheduling in [40], whereas one can find suboptimal schedule by shortest path search algorithm. In [41], Furuta et al. proposed an integer linear programming formulation of the same TDMA scheduling problem, which can be a practical real-time solution for large-scale sensor networks. Goldsmith et al. have also provided very elegant ways to build energy optimal schedule in [42], where the authors analyzed the tradeoff between data collection time (i.e., the schedule length) and energy consumption and proposed a scheduling algorithm which is proven to be energy optimal. Energy optimality is achieved under the condition that sensor nodes should wait for transmission completion of its child nodes. Hence, relay nodes have to only wake up at once, which minimizes the number of switches from sleep mode to active mode. Minimization of this number is crucially important, since (1) transient mode consumes extra energy; (2) guard time affects overhead. Hence, if a node waits for the transmission of its children then it is possible to aggregate the received packets. Thus, no extra energy is consumed by switches from sleep mode to active mode, which also minimizes the listening duration because of guard time. On the other hand, Varaiya et al. minimize schedule length (i.e., the data collection time) with centralized and distributed algorithms, respectively, in [43]. A previous work of Varaiya called PEDAMACS, can be found in [44]. Moreover the TreeMAC protocol [45] arranges schedule in order to maximize the throughput of the network.

In this paper, we define the scheduling problem such as Varaiya did in [43], thus the optimal schedule depends on the number of generated packets and the data collection tree. However, most of the solution only optimizes schedules in terms of minimizing either the schedule length or energy consumption (see Figure 2). We consider both performance metrics in order to strike an optimal trade-off between longevity and QoS. To demonstrate this point, in Figure 2 the energy and the schedule length minimal solutions are presented: (i) scheduling of Goldsmith et al. [42] does not exploit time slots for a given node until each packet arrived from its children; (ii) scheduling of [43] exploits time slots in a parallel way to minimize schedule length, but it suffers from significant energy overhead arisen from the extra on/off switches. In most health monitoring application a given data collection time is given, but energy minimization is still the primary objective. This paper presents a fast data gathering (comparable to [43]) with minimal energy consumption (comparable to [42]). This improvement is achieved from reducing the number of wake-ups using the information of application and routing layer.

3. The System Model

In this section, we present the system model concerning (i) the basic data collection framework; (ii) the energy model; (iii) the network model; (iv) the interference model.

3.1. The Data Collection Framework. In the case of wireless body-area monitoring network, sensor nodes worn on or implanted in the body continuously gather information about the monitored patient's physiological condition, such as electrocardiogram (ECG), pulse, blood oxygen saturation, skin temperature, blood pressure, and hemodynamic calculator [46]. These important measurements are stored in the local memory buffer of the node until the data is downloaded to the BS, when the patient arrives at a data collection point in a hospital. In this case it is imperative that the sensed data is acquired by the BS in an energy efficient way. This can be ensured by a proper scheduling algorithm. The sensor nodes located at different points on the human body transmit data in a multihop manner to the BS. Note, that both on-body and out-body sensors can be used to

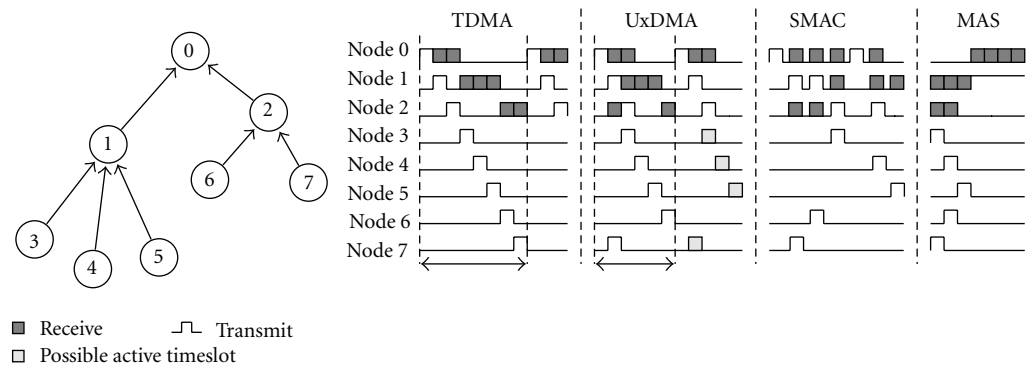


Figure 1: Traditional scheduling solutions of different MAC protocols. The possible active time slot indicates the fact that the node has the right to send (having the time slot) but it has no packet to transmit.

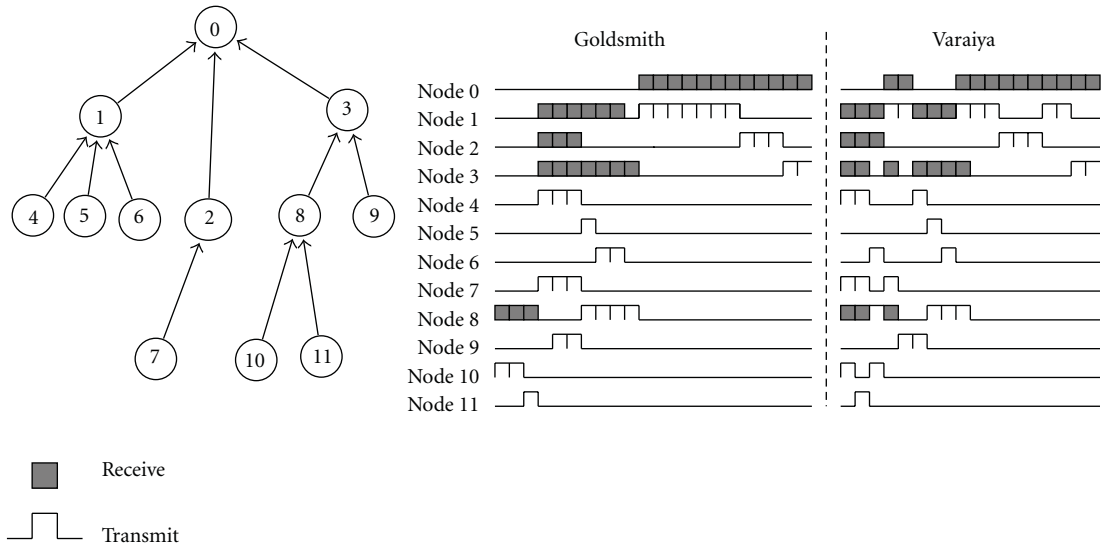


Figure 2: Recent scheduling solutions: Goldsmith et al. optimizes for energy consumption, Varaiya et al. optimizes for latency and throughput.

relay packets. In order to achieve efficient data transfer only a few protocols are designed specifically for multihop wireless body area network (WBAN) communication. One concern is to minimize the thermal effects of the implanted devices by balancing the traffic over a tree-like topology network.

In this paper, we are interested in developing an optimal scheduling for downloading the data from the low-power body sensors to BS. More specifically, the sensor nodes powered by limited capacity batteries, forward packets to other relay nodes by using a single-transmit and a single-receive antenna. The sensor nodes transmit the sensed information at a constant low power to their parents (i.e., parent node is the one which is selected from the routing table to relay the packet in the next hop) in a time-division manner based on the TDMA scheduling. The time axis is not divided into periodical frame periods such as in the case of [43] illustrated in Figure 2. Instead of using fixed frame size, a node is scheduled based on the routing condition and on the traffic requirements of the applications. For example, if a node has more tasks (sensed information and received packets to relay), then it wakes up more frequently. Based on this information each active time slot with a length of T_s is known to the sensor node. According to our novel packet scheduling scheme the BS decides on the time instant when a given sensor node utilizes the channel for T_s length of time. The proposed scheduling is made available to the nodes by using traditional dissemination protocols, which is detailed in Section 6. In the next subsection the energy model is presented which is based on the specification of Mica2 node with CC1100 transceiver.

3.2. The Energy Model. The target platform is the Berkeley Mica2 node [31] which is one of the most widely used WSN platforms. The platform has an 8 MHz processor, 4 kB of RAM, 128 kB of flash memory, and a 433 MHz wireless radio transceiver. The transfer rate is 38.4 kbps and it is powered by two AA batteries. Parameters related to energy use are listed in Table 1.

Table 1: Time, power and energy consumptions of a Mica2 node.

Operation process	Time	Power cons.	Energy cons.
Sleep	—	90 μW	—
Radio initialization	0.35 ms	18 mW	6.3 μJ
Turn on Radio	1.5 ms	3 mW	4.5 μJ
Switch to RX/TX state	0.25 ms	45 mW	11.25 μJ
Receive 1 byte	0.416 ms	45 mW	18.72 μJ
Transmit 1 byte	0.416 ms	60 mW	24.92 μJ

The slot size has to be greater than the length of packet times the duration needed to transmit/receive a byte plus the guard time. Formally we use the following notation:

$$T_s > L_p T_c, \tag{1}$$

where L_p is the number of bytes in a packet (with headers) and T_c is the time to transmit/receive a byte. Because of clock drift the receiver must wake up earlier with the so-called guard time, T_g. Hence,

$$T_s = T_g + L_p T_c. \tag{2}$$

Assuming that the default packet size is 28 bytes ($L_p = 28$) and the activation time of a sensor is 22 μJ (see Table 1), the following energy consumptions per slot can be derived for the receiver node:

$$E_{\mathrm{RX}} = \left(22 + 18.72\left(T_g + L_p\right)\right) \quad \mu\mathrm{J}, \tag{3}$$

and for the transmitter node

$$E_{\mathrm{TX}} = \left(22 + 24.92 L_p\right) \quad \mu\mathrm{J}. \tag{4}$$

During the inactive period, no data is exchanged between the sensor nodes. To efficiently save the energy, each sensor node is active during a time slot only when it has packet(s) to send or to receive, otherwise, it is in the sleep state and consumes $E_s\ [\mu J] = 90 \cdot T_s$. For these reasons, if a node has to send Y packets and receive $Y - X$ packets, then the following overall energy is consumed:

$$E = Y \cdot E_{\mathrm{TX}} + (Y - X) \cdot E_{\mathrm{RX}} + (K - 2Y - X)E_s, \tag{5}$$

where E is the total energy of the battery, hence K is the maximum lifespan of the network. One can see that the lifespan is basically determined by the traffic load X. In the next subsection the network model is introduced.

3.3. The WSN Model. We assume that the WSN has N static sensor nodes, which are able to transmit packets by using single omnidirectional antennas, and there exists a BS node to collect the data from the sensor nodes. These sensor nodes do not only send their own sensed data but they relay packets generated by other sensor nodes, as well. Each node has a local memory buffer, where the generated or received packets are temporarily stored. The WSN can be represented by a graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ denotes the set of nodes, and E denotes the set of edges referred to the communication links. If $\{v_i, v_j\} \subseteq V$, the edge $e = (v_i, v_j) \in E$ if and only if v_j is located within the transmission range of v_i. Let us define the *adjacency matrix* $\mathbf{A}$ of graph G (the size of matrix $\mathbf{A}$ is $N \times N$):

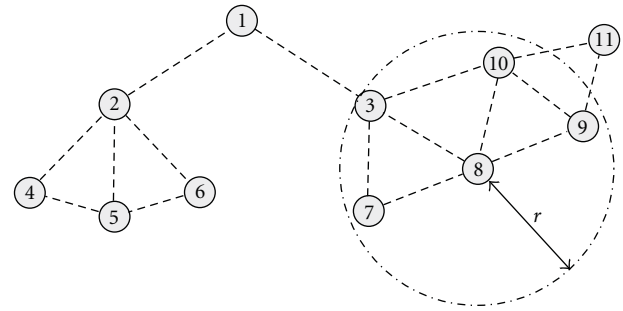


Figure 3: Nodes can only communicate in limited range r.

$$\mathbf{A} = \left[a_{ij}\right]_{i,j=1,\ldots,N}, \tag{6}$$

where

$$a_{ij} = 1, \tag{7}$$

if $e = (v_i, v_j) \in E$. These assumptions entail that all the sensor nodes have the same communication range r, the communication links are symmetrical, and no lossy links are assumed. Hence $\mathbf{A}$ is a symmetric matrix and represents the neighborhood information. This neighborhood information can be constructed either by measurements or by using general fading models.

Figure 3 shows a network example, where direct links only exist in limited range. Adjacency matrix of the example in Figure 3 is the following:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}. \tag{8}$$

To collect data a tree is constructed (which the most popular data gathering topology [43, 45]). The data gathering tree is routed at a BS node and each intermediate node collects the data from its children nodes and then forwards the data to its parent node. The routing tree is constructed by running the distributed Bellman-Ford algorithm (such as ESR, DSR). In Figure 4 a typical routing tree is illustrated.

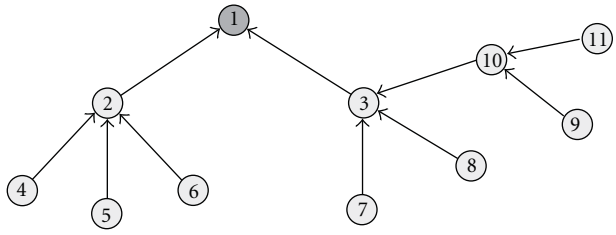


Figure 4: Example of a data collection tree. Nodes forwards the data to its parent node toward BS.

A static routing tree can be represented by a *routing matrix* with size $N \times N$:

$$\mathbf{R} = \left[r_{ij}\right]_{i,j=1,\ldots,N}, \tag{9}$$

where

$$r_{ij} = 1, \tag{10}$$

if node v_j is the parent of node v_i (i.e., if node v_i sends a packet, it sends to node v_j). The routing matrix of topology presented in Figure 4 is given as follows:

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{11}$$

Note that multiple packet transmissions are permitted, however, packet collision or interference has to be avoided.

3.4. The Interference Model. Due to the broadcast nature of the wireless medium, interference or collisions may occur among the nodes within the same transmission range. Two types of interference exist: primary interference and secondary interference [47]. Typical examples of primary interference is sending and receiving or receiving from two different transmitters at the same time slot. Secondary interference occurs when there are two parallel transmissions from different senders to different nodes being in the same collision domain (i.e., one of the transmission disturbs the signal at the receiver of the other transmission). Primary and secondary interferences are to be avoided in our approach. We use the protocol interference model [48] in this paper. In the protocol model, each node v_i has a fixed transmission range d and an interference range D, where $D \geq d$. For the sake of simplicity let us assume that $D = d$. In this case, according to the adjacency matrix, we define the *interference matrix* as follows:

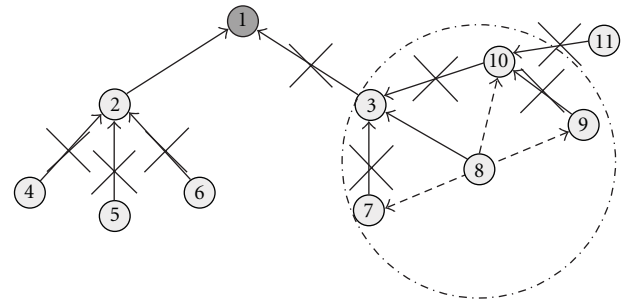


Figure 5: Collision situations in WSN. Only a few parallel transmission is possible while node 8 is sending a packet.

$$\mathbf{F} = \mathbf{A}^2 + \mathbf{A} - \mathrm{diag}(\mathbf{A}^2 + \mathbf{A}). \tag{12}$$

This construction is motivated by the fact that links are defined as follows:

$$\mathbf{F} = \left[f_{ij}\right]_{i,j=1,\ldots,N}, \tag{13}$$

where

$$f_{ij} = 1, \tag{14}$$

if node v_j is not allowed to transmit parallel with node v_i.

Figure 5 illustrates the interference domain of a given node. As summary interference can occur during transmission of node 8 (Figure 5) because of the following reasons:

(i) the receiver node (3) tries to send at the same time as node 8;

(ii) a node (e.g., one of 7 or 10) sends packet to the receiver of node 3;

(iii) a node in the collision domain of the receiver (e.g., 7, 9 or 10) sends packet at the same time, effecting collision at node 3;

(iv) the sender node disturbs other receivers' collision domain, when another node (e.g., 11) is transmitting.

Note that we make the diagonal element of matrix **F** zeros to represent that a node does not conflict with itself. We also note that matrix **F** symmetric, because of the symmetric feature of adjacency matrix. To give an

example the interference matrix of topology of Figure 3 is the following:

$$\mathbf{F} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{15}$$

Let us analyze the 8th row vector:

$$\mathbf{F}_8 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}. \tag{16}$$

If the vector contains "1" in a given position then the corresponding node is not allowed to send packet at the same time as node 8 (e.g., 1, 3, 7, 9, 10, 11).

4. The Scheduling Problem

In this section, we formulate the scheduling problem of MAS. In order to do this we need a data structure to represent a schedule.

4.1. Representation of a Schedule. The scheduling is represented by a binary matrix $\mathbf{C}$ called *scheduling matrix*, where

$$\mathbf{C} \in \{0, 1\}_{N \times L}. \tag{17}$$

The number of rows N is equivalent with the number of nodes in the network and L represents the number of time slots during which the data collection must be completed. The element $c_{jk} = 1$ if node v_j sends a packet to the receiver defined by the routing tree at time instant k, and 0 otherwise.

A possible schedule is given as follows:

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{18}$$

This particular matrix represents a schedule, where the different rows of the matrix define the TX schedule of the radios of the different nodes. Note that the RX schedule is uniquely determined by the scheduling matrix because of the tree topology of the data collection. If a node should send a packet according to the TX schedule then the first packet of the buffer is transmitted. For example, let us analyze the row vector related to node v_4:

$$\mathbf{c}_4 = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{19}$$

Based on this specific schedule, node v_4 transmits packet at time instants 1, 2, and 4 to its parent, where the parent is determined by routing matrix $\mathbf{R}$.

Returning to the challenge, we are seeking an optimal scheduling matrix $\mathbf{C}_{\text{opt}}$ of type $N \times L$, which guarantees interference free transmissions, minimizes the idle wake-ups and the energy consumption. In the case of exhaustive search the number of steps to obtain the solution is $\mathcal{O}(2^{NL})$. Thus, the endeavour is to find a feasible and efficient solution obtained in a much shorter time. To accomplish this objective, first we map the problem and the constraints into an objective function. Having the scheduling problem mapped into a quadratic objective function will let us use fast algorithms developed for quadratic optimization to solve the scheduling problem realtime.

Note that if we choose $L = K$ from (5), we are seeking schedule which is valid for the whole lifespan of the network. If L is too large then the whole problem is broken into smaller sized scheduling problems by choosing the size $L(p)$ as follows:

$$L = \sum_{p=1}^{P} L(p), \tag{20}$$

where $L(p)$ is the size of the pth smaller timescale problem.

Based on the discussion above, a WSN network is defined by its adjacency $\mathbf{A}_{N \times N}$, routing $\mathbf{R}_{N \times N}$, and interference $\mathbf{F}_{N \times N}$ matrices. Furthermore let matrix $\mathbf{C}_{N \times L(p)}$ denote the scheduling. In this and following subsections we look for a scheduling matrix, which guarantees feasibility and efficiency in terms of minimizing the following heuristic cost function and penalty functions.

4.2. The Objective Function. In this section we develop the objective function the optimization of which will guarantee the energy efficiency in terms of minimizing the RX/TX switches.

Since activation of the radio needs energy, it is worth selecting a schedule, in which the slots requiring radio activities (sending or receiving) follow each other to avoid frequent activations and deactivations. The following function rewards the situation when the receiver remains awake after an active slot, so the number of on/off switches is minimal by fulfilling the following expression:

$$\arg\min_{\mathbf{C}} \sum_{j=1}^{N} \sum_{k=1}^{L(p)-1} \Big(c_j(k) + c_j(k+1) - 2c_j(k)c_j(k+1) \Big). \tag{21}$$

4.3. Penalty Functions. The following penalty functions are developed to guarantee the feasibility and validity of the scheduling matrix $\mathbf{C}$.

```
Require: x, R, L(p)
    s, z ← x
    for k ← 1 to L(p) do
        z ← z · R
        s ← s + z
end for
```

Algorithm 1: Calculating the traffic vector.

4.3.1. Minimizing the Interference. In wireless networks minimizing the interference is equivalent with minimizing the number of collisions. Formally it means the minimization of

$$\min_{\mathbf{C}} \sum_{k=1}^{L(p)} \mathbf{C}(k)\mathbf{F}\mathbf{C}^T(k), \tag{22}$$

where $\mathbf{C}(k)$ denotes the kth column of scheduling matrix $\mathbf{C}$, the schedule corresponding to the time instant k. The quadratic function $\mathbf{C}(k)\mathbf{F}\mathbf{C}^T(k)$ gives the number of conflicts at time slot k. If this expression is zero then there is no packet which has collided.

4.3.2. Fulfilling the Transmission Requirements. Let us assume that the number of packets generated at node v_j is denoted by $x_j : x_j \geq 0$ and amount of packets must then be sent during $L(p)$ time slots to the BS from node v_j. Let us denote the number of transmissions needed (sending and forwarding) at node v_j as $s_j \geq x_j$, where s_j includes not only the sensed data packets but the packets to relay, as well. It is easy to see, that s_j at each node $v_j \in V$ can be easily calculated by Algorithm 1. This algorithm evaluates the traffic state s_j for all j runs from 1 to N.

Therefore s_j has to be equal with the sum of jth row in the scheduling matrix to perform all tasks provided by node v_j. Formally,

$$\forall v_j \in V : \sum_{k=1}^{L(p)} \mathbf{c}_j(k) = s_j. \tag{23}$$

This equation can be rewritten as the following minimization problem:

$$\min_{\mathbf{C}} \sum_{j=1}^{N} \left| s_j - \sum_{k=1}^{L(p)} \mathbf{c}_j(k) \right|. \tag{24}$$

Instead of using the absolute value let use minimize the following quadratic form:

$$\min_{\mathbf{C}} \sum_{j=1}^{N} \left(s_j - \sum_{k=1}^{L(p)} \mathbf{c}_j(k) \right)^2. \tag{25}$$

Note, that if $L(p)$ is too short then the minimum of (25) can be larger than zero, meaning that the remaining part of the traffic has to be forwarded in the next $L(p+1)$ session (e.g., as new task).

4.3.3. Minimizing the Remaining Number of Packets in the Network. The traffic generated at each sensor node (s_j) has to be collected by the BS for all j. It entails that the input and the output traffic in the network have to be the same. A scheduling vector of time instant k is denoted by $\mathbf{C}(k)$ which is the kth column vector of the scheduling matrix. Let us denote the buffer state vector in time instant k with $\mathbf{z}(k)$. Note, that multiplying the kth schedule vector and the routing matrix, the buffer state vector at the next time instant can be computed as

$$\mathbf{z}(k+1) = \mathbf{z}(k) + \mathbf{C}^T(k)\mathbf{R}. \tag{26}$$

The received vector $\mathbf{r}$ at time instant k represents the number of incoming packets and it is $\mathbf{z}(k+1) - \mathbf{z}(k) = \mathbf{C}^T(k)\mathbf{R}$. It is easy to see, that the number of incoming packets at time instant k at node v_i is

$$\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i). \tag{27}$$

The number of packets sent by a node v_i is

$$\sum_{k=1}^{L(p)} \mathbf{c}_i(k). \tag{28}$$

The number of incoming packets must equal the number of outgoing packet for each node, if

$$\forall v_i \in V : x_i + \left(\sum_{k=1}^{L(p)} \sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i) \right) = \sum_{k=1}^{L(p)} \mathbf{c}_i(k). \tag{29}$$

This can be reformulated as the following optimization task:

$$\min_{\mathbf{C}} \sum_{i=1}^{N} \left(x_i + \left(\sum_{k=1}^{L(p)} \sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i) \right) - \left(\sum_{k=1}^{L(p)} \mathbf{c}_i(k) \right) \right)^2. \tag{30}$$

4.3.4. Minimizing the Number of Idle Wake-Ups. The previous conditions guarantee that the information generated at sensor nodes arrive at the BS without interference. The penalty function developed in this subsection is to ensure that relay nodes should only wake up if there are packets to forward. This can be guaranteed if

$$\forall v_i \in V, \quad \forall l = 1 \cdots L(p),$$
$$\sum_{k=1}^{l} \left(\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i) \right) = \sum_{k=1}^{l} \mathbf{c}_i(k) \tag{31}$$

holds for each time instant. Formula (31) can be rewritten as follows:

$$\forall l = 1 \cdots L(p) :$$
$$\min_{\mathbf{C}} \sum_{i=1}^{N} \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i) \right) - \sum_{k=1}^{l} \mathbf{c}_i(k) \right)^2. \tag{32}$$

4.4. The Overall Objective Function. If we could minimize the previously defined objective function and four penalty functions simultaneously, then it will yield the scheduling matrix having the following properties:

(i) the schedule length is L;

(ii) all the sensed data packets x_j at each node v_j and incoming packet from other nodes are forwarded toward the BS;

(iii) schedules do not interfere with each other;

(iv) a node only wakes up if there is packet to send.

Therefore it is expected that the algorithm minimizing these objectives by minimizing the objective function (21) and the penalty functions (22), (25), (30), (32), can construct feasible and energy-efficient scheduling matrix **C**. In the next section we transform these optimization tasks into a single quadratic optimization by combining the objective function in an additive manner.

5. Quadratic Programming for Scheduling

Since the number of binary matrices is exponential with the number of nodes multiplied by the length of the schedule, exhaustive search with complexity

$$\mathcal{O}\left(2^{N\cdot L(p)}\right) \tag{33}$$

cannot be applied to solve the optimizations derived in the previous section. Thus, our goal is to develop a polynomial time solution by QP as it has been presented in [49–51]. In order to develop this solution let us first summarize the general background of QP.

5.1. Quadratic Programming. Let us assume that matrix **W** is a symmetric matrix of type $n \times n$ and vector **b** is of length n. In QP we seek the optimal n dimensional vector **y** which minimizes the following quadratic function [52]:

$$f(y) = \frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} + \mathbf{b}^T\mathbf{y}, \tag{34}$$

subject to one or more constraints of the form of

$$\begin{aligned} \mathbf{A}\mathbf{y} &\le \mathbf{v}, \\ \mathbf{B}\mathbf{y} &= \mathbf{u}. \end{aligned} \tag{35}$$

If QP contains only linear equation constraints then it can be solved as presented in [53]. In other cases, if the matrix **W** is positive definite, then the function $f(y)$ is convex and the problem can be solved with the ellipsoid method [20]. When **W** is indefinite the problem is NP hard (for details see in [54]).

A frequently used powerful heuristic algorithm to solve QP is the Hopfield Neural Network (HNN). This neural network is described by the following state transition rule:

$$y_i(k+1) = \operatorname{sgn}\left(\sum_{j=1}^{N} \widehat{W}_{ij}\, y(k)_j - \hat{b}_i\right), \quad i = \operatorname{mod}_N k, \tag{36}$$

where

$$\begin{aligned} \mathbf{d} &= -\operatorname{diag}(\mathbf{W}), \\ \widehat{\mathbf{W}} &= -\mathbf{W} - \operatorname{diag}(\mathbf{d}), \\ \hat{\mathbf{b}} &= \mathbf{b} - \frac{1}{2}\mathbf{d}. \end{aligned} \tag{37}$$

Using the Lyapunov method, authors in [55] proved that HNN converges to its fix point, as a consequence HNN minimizes a quadratic Lyapunov function:

$$\mathcal{L}(\mathbf{y}) := -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \widehat{W}_{ij}\, y_i y_j + \sum_{i=1}^{N} y_i \hat{b}_i = -\frac{1}{2}\mathbf{y}^T\widehat{\mathbf{W}}\mathbf{y} + \hat{\mathbf{b}}^T\mathbf{y}. \tag{38}$$

Thus, HNN is able to solve combinatorial optimization problems in polynomial time under special conditions [56–58]. As a result, one can see that there are quite a number of algorithms which can provide efficient solution to QP. Therefore it motivates us to map the scheduling problem to QP, which is done in the next subsection.

5.2. The Scheduling Problem as QP. First the binary scheduling matrix **C** is mapped into a binary column vector **y** as follows:

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1L} \\ C_{21} & C_{22} & \cdots & C_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ C_{J1} & C_{J2} & \cdots & C_{JL} \end{pmatrix} \longrightarrow \tag{39}$$

$$\mathbf{y} = (C_{11}, C_{21}, \ldots, C_{J1}, C_{12}, \ldots, C_{J2}, C_{1L}, \ldots, C_{JL})^T.$$

The objective function and various penalty functions are transformed into quadratic forms in the next paragraph. Once we have mapped each penalty function into a corresponding quadratic form, then we add them up to form an overall QP which then represents the scheduling problem.

5.2.1. Minimizing the Number of Radio On/Off Switches. The following mapping has to be carried out in order to construct a QP for this objective function:

$$\begin{aligned} &\sum_{j=1}^{N}\sum_{k=1}^{L(p)-1}\Big(c_j(k) + c_j(k+1) - 2c_j(k)c_j(k+1)\Big) \\ &\qquad = \frac{1}{2}\mathbf{y}^T\mathbf{W}_O\mathbf{y} + \mathbf{b}_O^T\mathbf{y}. \end{aligned} \tag{40}$$

Similarly to the previous subsection this construction penalizes the wrong solutions. For example a schedule has to be penalized if a node does not wake up again to transmit or receive after a some preceding radio activity. On the other hand, the schedule is rewarded in the case of the continuous radio activity. The corresponding objective function is described by (21).

Having solved (40) one obtains the following $\mathbf{W}_O$ matrix and $\mathbf{b}_O$ vector as follows:

$$\mathbf{b}_O = (-0.5, -1, -1, \ldots, -1, -0.5)_{(NL(p)\times 1)}, \tag{41}$$

$$\mathbf{W}_O = -\begin{pmatrix} \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N} & \cdots & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ & & \ddots & \ddots & \ddots & \\ \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} \\ \mathbf{0}_{N\times N} & \cdots & \cdots & \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} & \mathbf{0}_{N\times N} \end{pmatrix}. \tag{42}$$

5.2.2. Handling the Interference Constraint. In order to obtain the parameter of QP over binary vector $\mathbf{y}$ for the interference constraint, the following equation will yield the corresponding $\mathbf{W}_A$ and $\mathbf{b}_A$:

$$\sum_{k=1}^{L(p)} \mathbf{C}(k)\mathbf{F}\mathbf{C}^T(k) = \frac{1}{2}\mathbf{y}^T\mathbf{W}_A\mathbf{y} + \mathbf{b}_A^T\mathbf{y}. \tag{43}$$

It is easy to see that

$$\mathbf{b}_A = \mathbf{0}_{NL(p)\times 1}, \tag{44}$$

$$\mathbf{W}_A = 2\begin{pmatrix} \mathbf{F}_{N\times N} & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N} & \mathbf{F}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N} & \cdots & \mathbf{F}_{N\times N} \end{pmatrix}, \tag{45}$$

where $\mathbf{F}$ is the interference matrix.

5.2.3. Handling the Transmission Requirements. To obtain the parameters of the corresponding QP the following equation has to be solved for $\mathbf{W}_B$ and $\mathbf{b}_B$:

$$\sum_{j=1}^{N}\left(s_j - \sum_{k=1}^{L(p)} \mathbf{c}_j(k)\right)^2 = \frac{1}{2}\mathbf{y}^T\mathbf{W}_B\mathbf{y} + \mathbf{b}_B^T\mathbf{y}, \tag{46}$$

where vector $\mathbf{y}$ denotes the quantity of packets to be sent in a given time interval $L(p)$. Let us rewrite the left side of (46) as

$$\begin{aligned} &\sum_{j=1}^{N}\left(s_j - \sum_{k=1}^{L(p)} \mathbf{c}_j(k)\right)^2 \\ &\quad= \sum_{j=1}^{N}\left(s_j^2 + \left(\sum_{k=1}^{L(p)} \mathbf{c}_j(k)\right)^2 - 2s_j\sum_{k=1}^{L(p)} \mathbf{c}_j(k)\right) \\ &\quad= \sum_{j=1}^{N}\left(s_j^2\right) + \sum_{j=1}^{N}\left(\sum_{k=1}^{L(p)} \mathbf{c}_j(k)\right)^2 - 2\sum_{j=1}^{N}\left(s_j\sum_{k=1}^{L(p)} \mathbf{c}_j(k)\right). \end{aligned} \tag{47}$$

Note, that $\sum_{j=1}^{N} s_j^2$ is constant, which can be neglected from the point of minimization. By solving (46) one obtains

$$\mathbf{b}_B = (\mathbf{s}, \mathbf{s}, \ldots, \mathbf{s})_{NL(p)\times 1},$$

$$\mathbf{W}_B = 2\begin{pmatrix} \mathbf{I}_{N\times N} & \mathbf{I}_{N\times N} & \cdots & \mathbf{I}_{N\times N} \\ \mathbf{I}_{N\times N} & \mathbf{I}_{N\times N} & \cdots & \mathbf{I}_{N\times N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{N\times N} & \mathbf{I}_{N\times N} & \cdots & \mathbf{I}_{N\times N} \end{pmatrix}, \tag{48}$$

where $\mathbf{I}_{N\times N}$ is the identity matrix.

5.2.4. Handling the Constraint on the Remaining Number of Packets in the Network. Since the incoming traffic must be the same as the outgoing traffic for each node, we have

$$\begin{aligned} &\sum_{i=1}^{N}\left(\left(\sum_{k=1}^{L(p)}\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i)\right) - \left(\sum_{k=1}^{L(p)} \mathbf{c}_i(k)\right)\right)^2 \\ &\quad= \frac{1}{2}\mathbf{y}^T\mathbf{W}_D\mathbf{y} + \mathbf{b}_D^T\mathbf{y}. \end{aligned} \tag{49}$$

Let us rewrite the left side of (49):

$$\begin{aligned} &\sum_{i=1}^{N}\left(\sum_{k=1}^{L(p)}\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i)\right)^2 + \sum_{i=1}^{N}\left(\sum_{k=1}^{L(p)} \mathbf{c}_i(k)\right)^2 \\ &\quad- 2\sum_{i=1}^{N}\left(\sum_{k=1}^{L(p)}\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i)\right)\left(\sum_{k=1}^{L(p)} \mathbf{c}_i(k)\right). \end{aligned} \tag{50}$$

Evaluating the formula above results in the following quadratic form:

$$\mathbf{b}_C = \mathbf{0}_{NL(p)\times 1},$$

$$\mathbf{W}_C = 2\begin{pmatrix} \mathbf{P}_{N\times N} & \mathbf{P}_{N\times N} & \cdots & \mathbf{P}_{N\times N} \\ \mathbf{P}_{N\times N} & \mathbf{P}_{N\times N} & \cdots & \mathbf{P}_{N\times N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{N\times N} & \mathbf{P}_{N\times N} & \cdots & \mathbf{P}_{N\times N} \end{pmatrix}, \tag{51}$$

where matrix $\mathbf{P}$ of type $N \times N$ can be constructed as

$$\mathbf{P} = \mathbf{D}_C + \mathbf{I}_{N\times N} - 2\mathbf{R}_{N\times N}. \tag{52}$$

Note, that $\mathbf{D}_C$ is defined in (53).

$$\mathbf{D}_C = \begin{pmatrix} \mathbf{R}_{1,1}^2 + \mathbf{R}_{1,2}^2 + \cdots + \mathbf{R}_{1,N}^2 & \mathbf{R}_{1,1}\mathbf{R}_{2,1} + \cdots + \mathbf{R}_{1,N}\mathbf{R}_{2,N} & \cdots & \mathbf{R}_{1,1}\mathbf{R}_{N,1} + \cdots + \mathbf{R}_{1,N}\mathbf{R}_{N,N} \\ \mathbf{R}_{1,1}\mathbf{R}_{2,1} + \mathbf{R}_{1,2}\mathbf{R}_{2,2} + \cdots + \mathbf{R}_{1,N}\mathbf{R}_{2,N} & \mathbf{R}_{2,1}^2 + \mathbf{R}_{2,2}^2 + \cdots + \mathbf{R}_{2,N}^2 & \cdots & \mathbf{R}_{2,1}\mathbf{R}_{N,1} + \cdots + \mathbf{R}_{2,N}\mathbf{R}_{N,N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{1,1}\mathbf{R}_{N,1} + \mathbf{R}_{1,2}\mathbf{R}_{N,2} + \cdots + \mathbf{R}_{1,N}\mathbf{R}_{N,N} & \mathbf{R}_{2,1}\mathbf{R}_{N,1} + \mathbf{R}_{2,2}\mathbf{R}_{N,2} + \cdots + \mathbf{R}_{2,N}\mathbf{R}_{N,N} & \cdots & \mathbf{R}_{N,1}^2 + \mathbf{R}_{N,2}^2 + \cdots + \mathbf{R}_{N,N}^2 \end{pmatrix}. \tag{53}$$

5.2.5. Handling the Constraint on Minimizing the Number of Idle Wake-Ups. In order to map this constraint into a quadratic form the following transformation has to be carried out:

$$\forall l = 1 \cdots L(p) : \sum_{i=1}^{N} \left(\sum_{k=1}^{l} \left(\sum_{j=1}^{N} \mathbf{c}_j(k)\mathbf{R}(j,i) \right) - \sum_{k=1}^{l} \mathbf{c}_i(k) \right)^2 = \frac{1}{2}\mathbf{y}^T\mathbf{W}_D\mathbf{y} + \mathbf{b}_D^T\mathbf{y}. \tag{54}$$

In this case we penalize the wrong solutions. In order to solve this equation first let us analyze the traffic vector (i.e., the number of packets waiting in the buffers) at the different time instances. Obviously the traffic vector is $\mathbf{z}(0) = \mathbf{x}$ at time instant 0. At time instant 1 the traffic vector updated as follows:

$$\mathbf{z}(1) = \mathbf{x} - \mathbf{c}(1) + \mathbf{c}^T(1)\mathbf{R}. \tag{55}$$

Note, that it is only true if $\mathbf{c}(1)$ contains only "1"-s, if there is at least one packet to send. This formula can be evaluated recursively and the traffic vector can be calculated for every time instance. It is easy to see that $\mathbf{c}(2)$ has to be penalized if it contains "1" at a location where there is no packets in the buffer. Thus the penalty

$$(\mathbf{1} - \mathbf{z}(1))\mathbf{c}(2) \tag{56}$$

is positive if there is no packet at the buffer, but node (2) is waken up. On the other hand, it is negative if there is more than one packet in the buffer. Substituting (55) into (56) we get

$$\begin{aligned} &\left(\mathbf{1} - \mathbf{x} - \mathbf{c}(1) + \mathbf{c}(1)^T\mathbf{R}\right)\mathbf{c}(2) \\ &\quad = \mathbf{c}(1)(\mathbf{I} - \mathbf{R})\mathbf{c}(2) - (\mathbf{x} - 1)\mathbf{c}(2). \end{aligned} \tag{57}$$

Following this method recursively, we get the matrix of QP defined as

$$\mathbf{b}_D = (\mathbf{x} - 1, \mathbf{x} - 1, \ldots, \mathbf{x} - 1), \tag{58}$$

$$\mathbf{W}_D = 2\left(\mathbf{D}_D \cdot \mathbf{D}_D^T\right), \tag{59}$$

where matrix $\mathbf{D}_D$ of type $NL(p) \times NL(p)$ can be calculated as

$$\mathbf{D}_D = \begin{pmatrix} \mathbf{I}_{N\times N} - \mathbf{R} & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N} & \mathbf{I}_{N\times N} - \mathbf{R} & \cdots & \mathbf{0}_{N\times N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N} & \cdots & \mathbf{I}_{N\times N} - \mathbf{R} \end{pmatrix}. \tag{60}$$

5.3. Constructing the Overall QP for the Scheduling Problem. Based on the previously defined quadratic forms belonging to the different objective functions we combine them into an overall objective function as follows:

$$\min \frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} + \mathbf{b}^T\mathbf{y}, \tag{61}$$

where

$$\begin{aligned} \mathbf{W} &= \mathbf{W}_O + \alpha\mathbf{W}_A + \beta\mathbf{W}_B + \gamma\mathbf{W}_C + \delta\mathbf{W}_D, \\ \mathbf{b} &= \mathbf{b}_O + \alpha\mathbf{b}_A + \beta\mathbf{b}_B + \gamma\mathbf{b}_C + \delta\mathbf{b}_D. \end{aligned} \tag{62}$$

Note that, the objectives can be controlled with heuristic constants α, β, γ, δ in order to strike a good tradeoff between the importance of different requirements. Having these quadratic form at hand now we can apply the HNN (described in Section 5.1) to provide an approximate solution to the QP in polynomial time.

In the next section simulation results illustrate the advantages of our novel proposed scheduling algorithm.

6. Performance Evaluation

In this section we evaluate the performance of the novel algorithm and compare it to the traditional solutions.

We use a variety of microbenchmark tests to show how MAS improves channel access functionality. These benchmarks show the throughput and energy consumption of MAS, TreeMAC, PEDAMACS, S-MAC, and UxDMA, protocols developed by Varaiya et al. and Goldsmith et al., respectively. We will demonstrate that there is a tradeoff among throughput, latency, and energy consumptions. The figures obtained from the microbenchmarks empirically characterize the performance of MAS and provides an insight of the expected behavior of the scheduling algorithms in TDMA systems. To illustrate the effectiveness of our centralized scheduling algorithm, we compare MAS to existing MAC protocols, especially with the S-MAC, TreeMAC,

protocols developed by Varaiya and Goldsmith and UxDMA-based TDMA protocol.

The protocol developed by Goldsmith et al. [42] minimizes this number and also minimizes the the energy consumption of the nodes in the network, but it provides less throughput. On the other hand the two protocols developed by Varaiya et al. [43] provide relatively high throughput and minimizes the length of schedule, however they are not energy efficient in terms of minimizing the RX/TX switches. We also implemented TreeMAC [45] in order to analyze its throughput, whereas the primary goal was to maximize capacity. Our proposed protocol can achieve high throughput being comparable to protocol developed by Varaiya, on the other hand it works in an energy-efficient manner which extends the lifespan of the corresponding WSN. As a result, we managed to develop an adjustable protocol which provides a good tradeoff between throughput and energy efficiency.

At the last subsection we will summarize the complexity of the algorithms.

6.1. The Simulation Method. Each protocol has been simulated on a large set of traffic and topology parameters, which will characterize the protocol performance empirically. The purpose of these Matlab-based microbenchmarks is to show how the investigated protocols perform with different network topologies.

We simulated TDMA-based protocol variants using time slot assignment method called UxDMA [33]. Three different versions of UxDMA time slot assignment have been implemented. The data payload is the default payload for TinyOS applications; however, the protocols send only packets with length specified by higher level services. When we performed our tests, we used the parameters of Mica2 wireless sensor node to perform our tests (Table 2). All tests performed in an 100 m times 100 m area with nodes with no line-of-sight communication to every other node. The topology of the simulated networks is static, and we used both homogeneous and heterogeneous initial packed load on the sensor nodes. (In case of homogeneous packet load one packet is generated on each sensor node to send. In case of heterogeneous packet loads the initial number of packets are generated in random fashion.)

The multihop data collection trees were obtained by the Bellman-Ford algorithm using hop count as link metric. To determine the number of radio switches of each protocol, we implemented counters in the simulation that kept track of how many times the various operations were performed.

In all cases, we measured the data throughput, schedule length, and number of radio switches of the network. In all tests "packet size" is referring to the size of the data payload without the header information. The overhead attributed to each protocol is shown in Table 2. Note that control packets in S-MAC are less (18 bytes) than in the case of MAS, however, the S-MAC overheads consume energy at every packets. On the other hand MAS produces control overhead to disseminate the actual schedule but it happens less frequently.

Table 2: Length of data and control packets used by protocols MAS, UxDMA, and S-MAC.

	MAS	UxDMA	S-MAC
Data Packet Length (bytes)			
Preamble	8	8	8
Synchronization	0	0	0
Header	6	6	6
Footer (CRC)	2	2	2
Data Length	26	26	26
Total	42	42	42
Control Packet Length (bytes)			
Preamble	8	0	8
Synchronization	0	0	0
Header	6	0	6
Footer (CRC)	2	0	2
Data Length	26	0	6
Total	42	0	18

The time synchronization of the nodes could be performed by FTSP protocol [59]. This protocol is fast enough and requires only one dissemination cycle by the BS. The dissemination of the schedule tables can be integrated into the FTSP method. The topology and interference discovery we used is similar as Varaiya et al. introduced in [43]. This protocol operates in two phases: (i) first all nodes determine their neighbors and broadcasts their information about the neighbors, (ii) this information reaches the BS then the BS determines the routing matrix, and interference matrix, topology matrix. The amount of packet which have to be scheduled may be included into the previous data messages, therefore with the discovery protocol the BS can acquire information about the number of packet to be scheduled.

In the simulation the synchronization packets, the network topology discovery packets, and schedule control packets of each protocol are neglected (i.e., each solution requires the same amount of packets to keep the nodes' synchronicity and to disseminate actual schedule).

6.2. Throughput. Throughput or channel utilization is one of the most important metrics for MAC protocols which illustrates protocol efficiency. High channel utilization is critical for delivering a large number of packets in a short amount of time. In sensor networks, the speed of data transfer can be important, for example, to detect emergency situations. MAS minimizes the time to send packets as opposed to TDMA systems with fix frame size. MAS also achieves lower contention than the contention-based protocols. In the test we analyze the performance over 100 different network topologies including 30 nodes and the BS collects total number of 30 packets generated from sensor nodes. Each node generates one packet in homogeneous scenario. The throughput of the network was taken as the average of the 100 different simulation results. The throughput achieved by the different protocols is shown in Figures 6 and 7. Whenever the initial packet load is heterogeneous the achieved throughput is better than in case

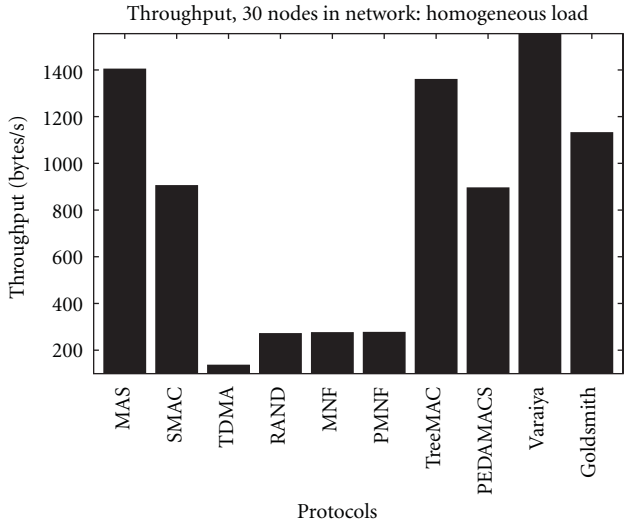


Figure 6: Throughput comparison of the simulated protocols in case of WSN with 30 nodes and homogeneous packet load.

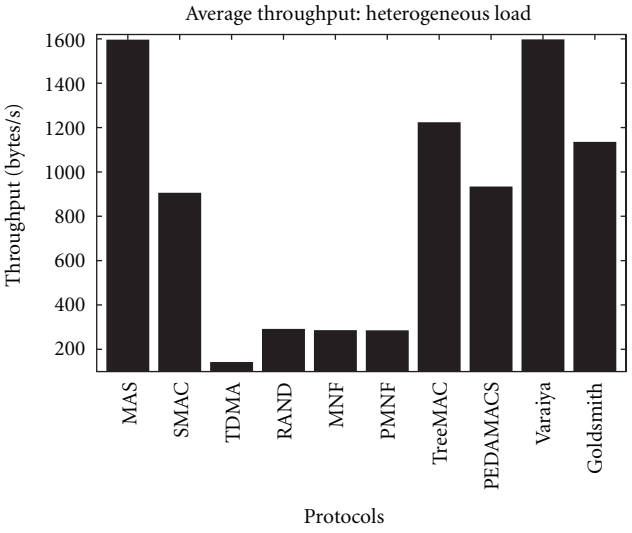


Figure 7: Throughput comparison of the simulated protocols in case of WSN with 30 nodes and heterogeneous packet load.

of homogeneous load. One may notice that the TreeMAC protocol falls back due to the enormous number of unused timeslots.

Varaiya's method gives the best solution as far as the throughput is concerned, however our solution only slightly differs from that that of Varaiya's. One can see that MAS aims to find the optimal solution which has a maximal throughput. In general, better throughput is accomplished with our novel scheduling algorithm: MAS outperforms S-MAC, PEDAMACS, Goldsmith's method, and all variant of the UxDMA-based TDMA protocols. Furthermore, the TreeMAC solution, which is also developed for good throughput, is also outperformed by MAS. S-MAC has a lower channel utilization performance only because of suffering from the overhead of RTS-CTS (Request To Send,

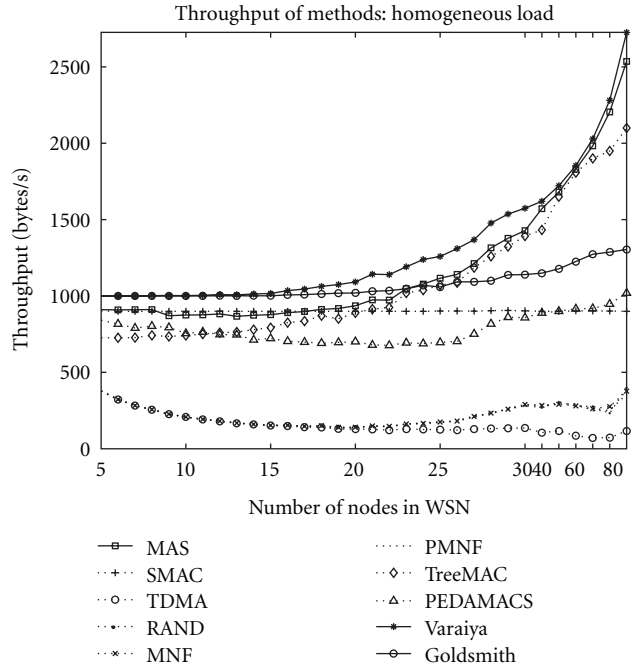


Figure 8: Throughput comparison of the simulated protocols in the case of different network size, using homogeneous packet load.

Clear To Send) exchanges. Instead of using control messages like RTS-CTS for hidden terminal support, MAS uses a control packet to inform the nodes about the optimal schedule, this control packets also used for synchronizing the nodes. In Figure 8, one can see as the number of nodes increases in the network the performance of MAS converges to performance of Varaiya's protocol. MAS can utilize several times more time slots, than the static frame-based TDMA solutions. The tests demonstrated that channel utilization of MAS is higher than the traditional MAC solutions and also higher from novel distributed solutions.

We designed MAS to operate in an energy-efficient manner by minimizing the number of radio on/off and the number TX/RX switches. In the previous subsection we demonstrated that the throughput of MAS is high and now we show that the novel method is also energy efficient. Low duty cycle applications have extremely low network throughput, and vice versa, high-throughput applications need a lot of energy. However, some applications, such as bulk data transfer also need protocols ensuring high throughput with energy efficiency. In this subsection we evaluate the number of radio switches of different protocols in different network scenarios. Figure 9 shows the number of radio switches of the different methods according to the network size.

Figure 10 summarizes the overall number of radio switch overhead of the network, whose result is simulated as the average of 100 different scenarios with 30 nodes and one packet on each node. The protocol designed by Goldsmith is the best from this point of view. Both the traditional TDMA and Varaiya's protocol has more energy overhead than Goldsmith's method.

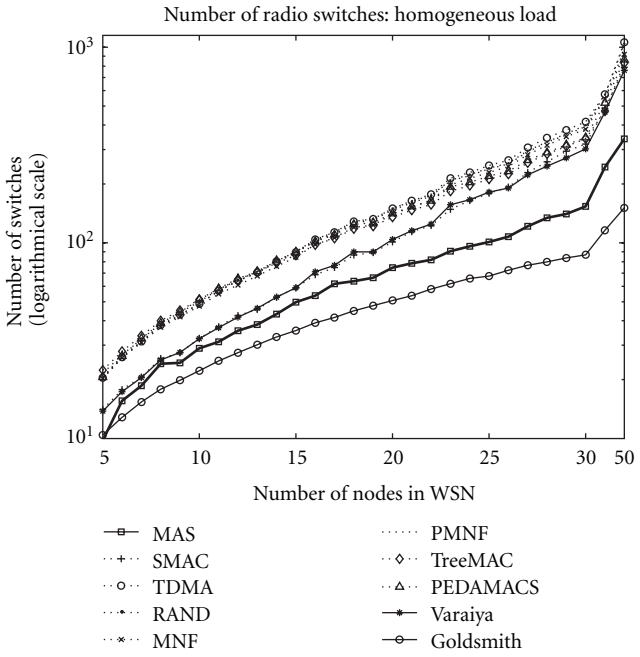


Figure 9: Number of radio switches in schedule of the simulated protocols in the case of different network size, using homogeneous packet load.

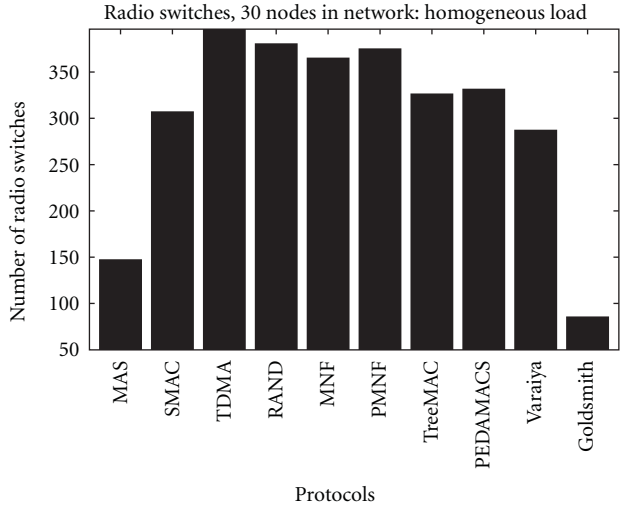


Figure 10: Average number of radio switches in a network containing 30 nodes, using homogeneous packet load.

In case of heterogeneous initial packet load the differences are even bigger. In Figure 11, one can see the number of radio switches of the different protocols. The scenario is the same: 30 nodes and random number of packets packets. The energy efficiency of the MAS protocol is between the methods developed by Varaiya and Goldsmith.

6.4. Energy Efficiency. When the MAC protocol is permitted to increase latency, we can reduce the duty cycle of the node and conserve energy. The latency is measured by the total length of schedule.

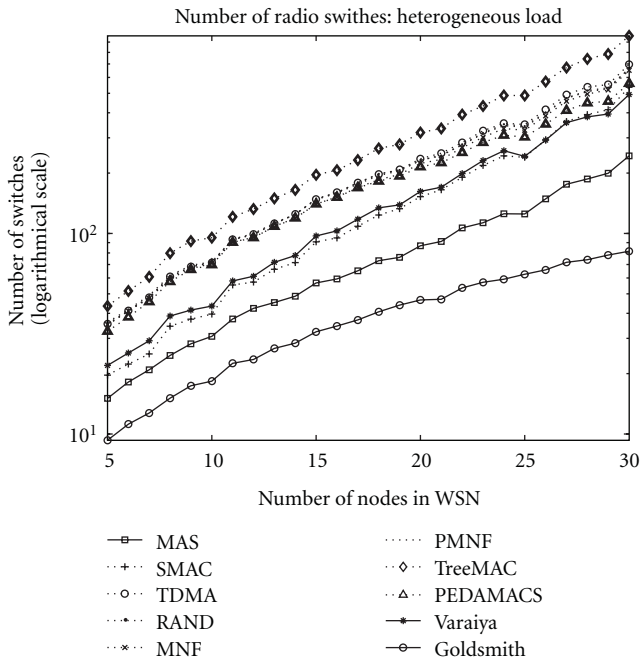


Figure 11: Number of radio switches in schedule of the simulated protocols in the case of different network size—heterogeneous initial packet load.

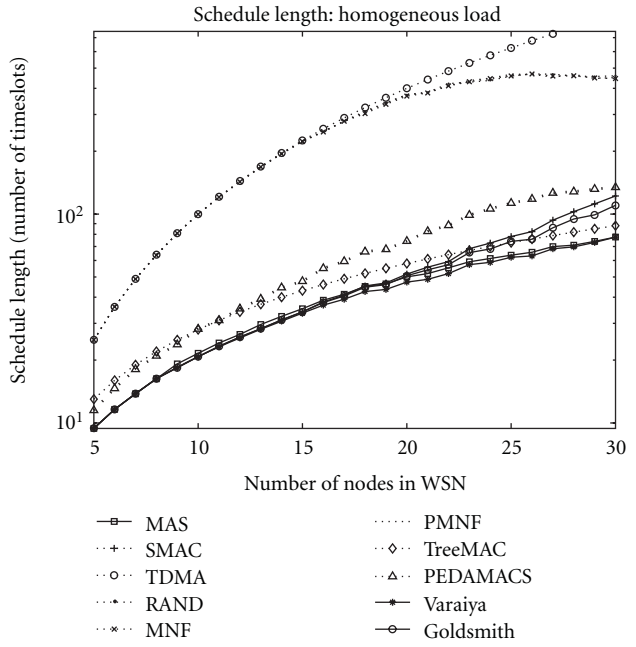


Figure 12: Schedule length comparison of the simulated protocols in the case of different network size—homogeneous packet load.

Result of test for evaluating end-to-end latency can be seen in Figure 12 for homogeneous packet load and in Figure 13 for heterogeneous load. Figure 14 summarizes schedule length differences amongst the tested protocols. MAS protocol has better values for heterogeneous packet load, outperforms the Goldsmith's method and performs as well as the Varaiya's method. Note that in

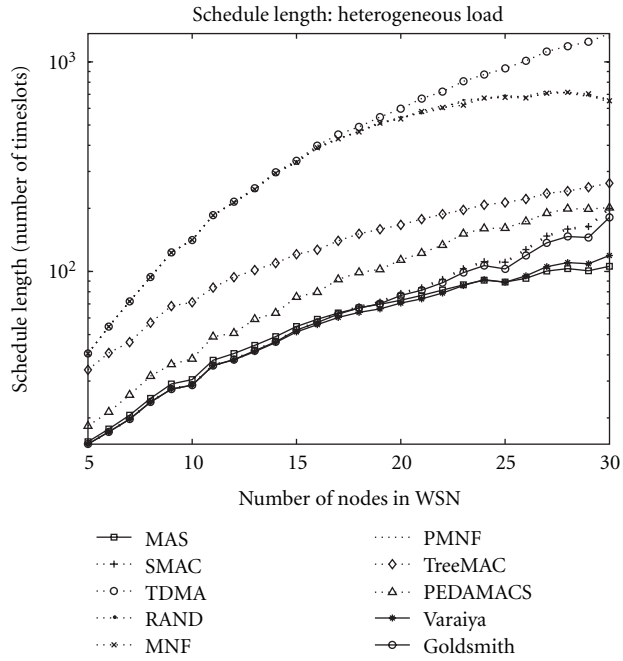


Figure 13: Schedule length comparison of the simulated protocols in the case of different network size—heterogeneous packet load.

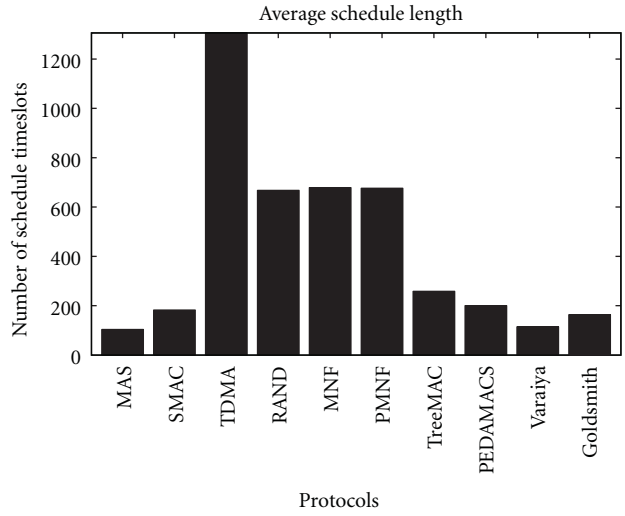


Figure 14: Schedule length comparison of the simulated protocols—WSN with 30 nodes.

the case of UxDMA, minimization of frame size is the primary objective in order to yield as low latency as possible. However, it is clear from figures that the end-to-end delay is much worse in the case of static frame sized TDMA protocol than in the case of MAS or S-MAC.

Based upon the previous test, we have a new protocol which achieves good throughput, minimizes the number of radio switches (and the consumed energy) and has minimal latency as well. It combines the good properties of several protocols.

Table 3: Algorithmic complexity of investigated protocols, where $d_{\max}$ denotes the maximal degree of nodes in the topology tree.

Method	Complexity
MAS	$\mathcal{O}(L^2 * N^2)$
SMAC	$\mathcal{O}(1)$
TDMA	$\mathcal{O}(1)$
RAND	$\mathcal{O}(N^2)$ (coloring)
MNF	$\mathcal{O}(N^2)$ (coloring)
PMNF	$\mathcal{O}(N^2)$ (coloring)
TreeMAC	$\mathcal{O}(N*$ Number of levels in topology tree)
PEDAMACS	$\mathcal{O}(L * N * d_{\max})$ + coloring
Varaiya	$\mathcal{O}(L * N * d_{\max})$ + coloring
Goldsmith	$\mathcal{O}(L * N * d_{\max})$

6.5. Complexity Analysis of the Algorithms. The complexity of the protocols is an important metrics in order to determine how fast we can achieve the scheduling and the corresponding timeslot assignment. Table 3 summarizes the complexity of the investigated protocols. All of tested protocols has polynomial or better complexity, however it is obvious that the novel MAS protocol may be slower than the others.

7. Conclusions

We have proposed a novel protocol to provide an optimal scheduling for data collection in Wireless Sensor Networks. The objective was to minimize the number of radio RX/TX switches in order to achieve minimum energy overhead. Other factors, such as (i) achieving interference free communication; (ii) minimizing the number of idle wake-ups; (iii) gathering the specified amount of packets within the given latency constraints, have been taken into account as penalties. Then optimal scheduling has been transformed into quadratic functions, which can then be solved heuristically in polynomial time by using the Hopfield Neural Network. Simulations have demonstrated that the novel MAS protocol minimizes energy consumption and delays. The performance evaluation has shown that the achieved end-to-end latency is as good as either the traditional contention or the traditional scheduled based solutions. On the other hand, MAS consumes less energy by minimizing the number of radio switches. Furthermore, MAS manages to achieve high channel utilization (comparable to Varaiya's method), with low energy consumption (comparable to Goldsmith's method).

References

[1] P. Corke, T. Wark, R. Jurdak, D. Moore, and P. Valencia, "Environmental wireless sensor networks," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1903–1917, 2010.

[2] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

[3] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications (WSNA '02)*, pp. 88–97, ACM, New York, NY, USA, 2002.

[4] D. Puccinelli and M. Haenggi, "Wireless sensor networks: applications and challenges of ubiquitous sensing," *IEEE Circuits and Systems Magazine*, vol. 5, no. 3, pp. 19–31, 2005.

[5] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

[6] A. Rogers, D. D. Corkill, and N. R. Jennings, "Agent technologies for sensor networks," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 13–17, 2009.

[7] T. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 2001.

[8] R. Jurdak, C. V. Lopes, and P. Baldi, "A survey classification and comparative analysis of medium access control protocols for ad hoc networks," *IEEE Communications Surveys and Tutorials*, vol. 6, pp. 2–16, 2004.

[9] J. Levendovszky, L. Tran-Thanh, G. Treplan, and G. Kiss, "Fading-aware reliable and energy efficient routing in wireless sensor networks," *Computer Communications*, vol. 33, 1, pp. S102–S109, 2010.

[10] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS '00)*, vol. 8, p. 223, IEEE Computer Society, Honolulu, Hawaii, 2000.

[11] S. Cui, R. Madan, A. J. Goldsmith, and S. Lall, "Joint routing, MAC, and link layer optimization in sensor networks with energy constraints," in *Proceedings of the IEEE International Conference on Communications (ICC '05)*, pp. 725–729, Seoul, Korea, 2005.

[12] I. Akyildiz and I. Kasimoglu, "Wireless sensor and actor networks: research challenges," *Ad Hoc Networks*, vol. 2, no. 4, pp. 351–367, 2004.

[13] K. H. Hui, D. Guo, R. Berry, and M. Haenggi, "Performance analysis of MAC protocols in wireless line networks using statistical mechanics," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, pp. 1315–1322, Monticello, Ill, USA, 2009.

[14] J. Polastre, J. Hui, P. Levis et al., "A unifying link abstraction for wireless sensor networks," in *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems (SenSys '05)*, pp. 76–89, ACM, New York, NY, USA, 2005.

[15] L. F. W. V. Hoesel and P. J. M. Havinga, "A lightweight medium access protocol for wireless sensor networks," Tech. Rep., 2004.

[16] D. Chen and P. K. Varshney, "On-demand geographic forwarding for data delivery in wireless sensor networks," *Computer Communications*, vol. 30, no. 14-15, pp. 2954–2967, 2007.

[17] M. Schuts, F. Zhu, F. Heidarian, and F. W. Vaandrager, "Modelling clock synchronization in the chess gMAC WSN protocol," in *Proceedings of the Workshop on Quantitative Formal Methods: Theory and Applications (QFM '09)*, vol. 13, Eindhoven, The Netherlands, 2010.

[18] I. Demirkol, C. Ersoy, and F. Alagoz, "MAC protocols for wireless sensor networks: a survey," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 115–121, 2006.

[19] M. Stemm, R. H. Katz, and Y. H. Katz, "Measuring and reducing energy consumption of network interfaces in hand-held devices," *IEICE Transactions on Communications*, vol. 80, no. 8, pp. 1125–1131, 1997.

[20] Z. Q. Luo, W. K. Ma, A.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 20–34, 2010.

[21] R. Madan, S. Cui, S. Lall, and A. J. Goldsmith, "Modeling and optimization of transmission schemes in energy-constrained wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 15, no. 6, pp. 1359–1372, 2007.

[22] N. Abramson, "The aloha system—another alternative for computer communications," in *Proceedings of the Fall Joint Computer Conference, AFIPS Conference*, ACM, Honolulu, Hawaii, 1970.

[23] A. Cawood and R. Wolhuter, "Comparison and optimisation of CSMA/CA back-off distribution functions for a low earth orbit satellite link," in *Proceedings of the IEEE AFRICON 2007*, pp. 1–8, Windhoek, South Africa, 2007.

[24] J. Polastre, J. Hill, and D. Culler, "Versatile low power media access for wireless sensor networks," in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems (SenSys '04)*, pp. 95–107, ACM, Baltimore, Md, USA, 2004.

[25] "Wireless lan medium access control (mac) and physical layer (phy) specifications," IEEE Standard 802.11, 1999.

[26] M. Buettner, G. V. Yee, E. Anderson, and R. Han, "X-MAC: a short preamble MAC protocol for duty-cycled wireless sensor networks," in *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems (SenSys '06)*, pp. 307–320, ACM, New York, NY, USA, 2006.

[27] S. Liu, K.-W. Fan, and P. Sinha, "CMAC: an energy-efficient MAC layer protocol using convergent packet forwarding for wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 5, no. 4, pp. 29–1, 29–34, 2009.

[28] Incel, Ghosh, and Krishnamachari, "Scheduling algorithms for tree based data collection in wireless sensor networks," in *Theoretical Aspects of Distributed Computing in Sensor Networks*, S. Nicoletseas and J. Rolim, Eds., Springer, 2010.

[29] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," in *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '02)*, pp. 1567–1576, New York, NY, USA, 2002.

[30] TOSWiki, "Tinyos introduction," 2010, http://docs.tinyos.net/tinywiki/index.php/Main_Page.

[31] CAPSIL, "Mica2 technical parameters," 2010, http://www.capsil.org/capsilwiki/index.php/MICA2.

[32] K. Sohrabi, J. Gao, V. Ailawadhi, and G. J. Pottie, "Protocols for self-organization of a wireless sensor network," *IEEE Personal Communications*, vol. 7, no. 5, pp. 16–27, 2000.

[33] S. Ramanathan, "A unified framework and algorithm for channel assignment in wireless networks," *Wireless Networks*, vol. 5, no. 2, pp. 81–94, 1999.

[34] I. Rhee, A. Warrier, J. Min, and L. Xu, "Drand: distributed randomized TDMA scheduling for wireless ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 10, pp. 1384–1396, 2009.

[35] I. Rhee, A. Warrier, M. Aia, and J. Min, "Z-MAC: a hybrid MAC for wireless sensor networks," in *Proceedings of the*

3rd International Conference on Embedded Networked Sensor Systems, pp. 90–101, ACM, New York, NY, USA, 2005.

[36] K.-P. Shih, C.-Y. Chang, Y.-D. Chen, and T.-H. Chuang, "Dynamic bandwidth allocation for QoS routing on TDMA-based mobile ad hoc networks," *Computer Communications*, vol. 29, no. 9, pp. 1316–1329, 2006.

[37] S. Cui, R. Madan, A. J. Goldsmith, and S. Lall, "Cross-layer energy and delay optimization in small-scale sensor networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 10, pp. 3688–3699, 2007.

[38] A. G. Ruzzelli, G. O'Hare, M. O'Grady, and R. Jurdak, "Merlin: cross-layer integration of MAC and routing for low duty-cycle sensor networks," *Ad Hoc Networks*, vol. 6, no. 8, pp. 1238–1257, 2007.

[39] R. Jurdak, P. Baldi, and C. V. Lopes, "Adaptive low power listening for wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 8, pp. 988–1004, 2007.

[40] T. Ukai, M. Sasaki, F. Ishizaki, and A. Suzuki, "A new approachfor scheduling problem in multi-hop sensor networks," in *Proceedings of the 9th International Symposium on Operations Research and Its Applications (ISORA '10)*, pp. 386–393, ORSC & APORC, Jiuzhaigou, China, 2010.

[41] T. Furuta, M. Sasaki, F. Ishizaki, T. Ukai, H. Miyazawa, and W. Koo, "New formulation for scheduling problem in multi-hop wireless sensor networks," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference (IWCMC '10)*, pp. 73–78, ACM, New York, NY, USA, 2010.

[42] S. Cui, R. Madan, A. Goldsmith, and S. Lall, "Energy-delay tradeoffs for data collection in TDMA-based sensor networks," in *Proceedings of the IEEE International Conference on Communications (ICC '05)*, pp. 3278–3284, Seoul, Korea, 2005.

[43] S. C. Ergen and P. Varaiya, "TDMA scheduling algorithms for wireless sensor networks," *Wireless Networks*, vol. 16, no. 4, pp. 985–997, 2010.

[44] S. Ergen and P. Varaiya, "Pedamacs: power efficient and delay aware medium access protocol for sensor networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 7, pp. 920–930, 2006.

[45] W.-Z. Song, R. Huang, B. Shirazi, and R. LaHusent, "TreeMAC: Localized TDMA MAC protocol for real-time high-data-rate sensor networks," in *Proceedings of the 7th Annual IEEE International Conference on Pervasive Computing and Communications (PerCom '09)*, pp. 1–10, 2009.

[46] D. Cypher, N. Chevrollier, N. Montavont, and N. Golmie, "Prevailing over wires in healthcare environments: benefits and challenges," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 56–63, 2006.

[47] S. Ramanathan and E. L. Lloyd, "Scheduling algorithms for multihop radio networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 2, pp. 166–177, 1993.

[48] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[49] J. Levendovszky, E. László, K. Tornai, and G. Treplán, "Optimal pricing based resource management," in *Proceedings of the International Conference on Operations Research*, p. 169, Germany, 2010.

[50] J. Levendovszky, A. Oláh, K. Tornai, and G. Treplán, "Novel load balancing algorithms ensuring uniform packet loss probabilities for WSN," in *Proceedings of the IEEE 73rd Vehicular Technology Conference*, Budapest, Hungary, 2011.

[51] E. László, K. Tornai, G. Treplán, and J. Levendovszky, "Novel load balancing scheduling algorithms for wireless sensor networks," in *Proceedings of the 4th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ '11)*, Budapest, Hungary, 2011.

[52] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, Berlin, Germany, 2006.

[53] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann, Berlin, Germany, 1988.

[54] S. Sahni, "Computationally related problems," *SIAM Journal on Computing*, vol. 3, no. 4, pp. 262–279, 1974.

[55] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 8, pp. 2554–2558, 1982.

[56] J. J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problems," *Biological Cybernetics*, vol. 55, no. 3, pp. 141–146, 1985.

[57] J. Mandziuk, "Solving the travelling salesman problem with a hopfield—type neural network," *Demonstratio Mathematica*, vol. 29, no. 1, pp. 219–231, 1996.

[58] J. Mandziuk and B. Macuk, "A neural network designed to solve the N-queens Problem," *Biological Cybernetics*, vol. 66, no. 4, pp. 375–379, 1992.

[59] M. Maróti, B. Kusy, G. Simon, and K. Lédeczi, "The flooding time synchronization protocol," in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems (SenSys '04)*, pp. 39–49, New York, NY, USA, 2004.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 939864, 11 pages
doi:10.1155/2011/939864

Research Article

Link Prediction and Route Selection Based on Channel State Detection in UASNs

Jian Chen, Yanyan Han, Deshi Li, and Jugen Nie

School of Electronic Information, Wuhan University, Wuhan 430079, China

Correspondence should be addressed to Yanyan Han, hanyan4981@163.com

Received 11 March 2011; Revised 31 May 2011; Accepted 6 July 2011

In Underwater Acoustic Sensor Networks (UASNs), data route is often disrupted by link interruption which will further lead to incorrect data transmission due to high propagation delay, Doppler effect, and the vulnerability of water environment in acoustic channel. So how to correctly transmit data when there are interrupted links on the data path is just the issue Delay Tolerant Networks (DTNs) aim to solve. In this paper, we propose a model to predict link interruption and route interruption in UASNs by the historical link information and channel state obtained by periodic detection. A method of decomposing and recomposing routes hop by hop in order to optimize route reselection is also presented. Moreover, we present a back-up route maintenance scheme to keep back-up routes with fresh information. In case of single route, we advance the idea to utilize the periodicity of environmental changes to help predict link interruption. In the simulation, we make comparisons on node energy consumption, end-to-end delay as well as bit error rate with and without link prediction. It can be derived that the network performance is significantly improved with our mechanism, so that our mechanism is effective and efficient while guaranteeing reliable data transmission.

1. Introduction

Wireless ad hoc network is a special kind of mobile communication networks with the characteristics of frequently changing network topology, multihop communication, and extremely limited energy [1]. In ad hoc networks, stable route is significant for reliable data transmission since it can decrease the calculation of route selecting, number of reply packets, alleviate energy loss of routers and shorten the packet propagation delay. The route stability relies on the stability of links in the route. However, links with higher stability can dramatically decrease the rerouting times and the probability of link interruption and consequently result in a long-term available route. Currently, the research in the area of wireless network is mainly based on the assumption: there exists at least one stable path between two ends. However, this assumption is not always satisfied [2]. For example, the wireless links between two different nodes could be broken as a result of natural environment, human interference, and nodes mobility. In this situation, how to make two ends restore normal communication to ensure real-time data collection and transmission is one of the challenging issues confronted by Delay Tolerant Network (DTN).

Currently, the study in DTN mainly focuses on mobile ad hoc network, wherein the topology changes strongly resulting from nodes roaming, and some links are unexpectedly broken in certain period of time. Moreover, DTN includes intermittently disconnected links with respect to the underwater acoustic network. And the model we propose here is applied in this environment.

Wireless sensor networks could be widely applied in environmental monitoring. But the intermittent link connection and topology and end-to-end variation in WSNs entails a huge challenge on how to guarantee the reliability of network operation and real-time data transmission. Compared with terrain sensor network, UASN has higher link interruption probability. The reasons of link interruption primarily include three aspects. (i) Network structure: node failures commonly exist due to energy exhaustion and topology variation that is caused by node floating. Besides, the fact the acoustic modem varies in its communication ranges has negative influence on the network topology generation. (ii) Underwater environment: ocean current, tide, and the sound-wave transmitted by mobile objects such as fish flock and ships passing by can disturb acoustic channel. (iii) Limitation of channel characteristics: underwater

acoustic channel has higher transmission delay due to low transmission rate. Moreover, Doppler effect and multipath effect will lead to distinct interference at the receiver that may produce incorrect data reception.

It is widely acknowledged that there are several disadvantages in underwater communication. First, the data storage requirement is relatively high when data need to be buffered for a long time before finally dropping. Secondly, the sink node will continually transmit inquiring messages if it does not receive data from nodes along the route, and nodes in the network will passively wait for the link recovery with full power, which both introduce high energy consumption. In order to address these problems, we put forward a model of underwater DTN link prediction and route selection based on channel state detection.

Specifically, we investigate how to detect acoustic channel and environment state effectively, then estimate future link availability according to detected information, and make corresponding decisions by estimation results. In our model, nodes can sleep for a certain time or choose a new route and retransmit data stored in the buffer. This can increase the probability of successful delivery while conserving energy of underwater nodes. Moreover, under our model, nodes can predict the duration of disconnection as well as connection efficiently and optimize the decision-making by leveraging the characteristics of underwater channel and periodicity of underwater environment.

This paper presents a link prediction model of underwater DTN; it considers the characteristics of acoustic channel in UASN and makes use of channel state information by periodic channel detection. Then the prediction results are applied to multiple-route switch and route optimization. Since this model utilizes the information of real-time acoustic channel and the historical records of nodes, it has great pertinence on networks deployed in underwater environment and can take effective measures for link interruption caused by environmental factors such as nodes floating and disturbances from fish flocks. We analyze and describe the process of decision making in cases of multiple routes and single route, respectively.

The rest of this paper is organized as follows. Section 2 includes a detailed survey of related work. Section 3 describes the whole model of this algorithm and its application background. Section 4 exhibits a specification of channel detection, prediction model, and decision making. Simulation results and analysis are presented in Section 5. Section 6 concludes the paper.

2. Related Work

The research on DTN is generally focused on network structure, routing mechanism, and reliability, and most of them are related to terrain mobile adhoc network. There are some literatures centering on algorithms of route selection and route maintenance. Wherein [3] presents that a distributed algorithm makes different numbers of copies of messages sent by nodes by flooding to ensure reliable data transmission even with some broken links. Packets are divided through priorities to determine the number of copies of packets, together with redundant information and resources reallocation. It can achieve that more important information will obtain more resources for less transmission delay, but there is no specification on how to handle route failure. In [4], a virtual network topology in mobile ad hoc network (MANET) is built using the statistical information collected by gateways to estimate future linking probability between neighbors. Unfortunately, this method uses gateway to make virtual topology and broadcast in the whole network, unsuitable for energy-limited UASNs. In [5], link stability is classified into different levels in terms of real-time measures such as signal strength. Route with the highest stability level will be chosen as main route. But it does not make prediction or analysis of future link state. So if the selected route was actually broken, this route would be invalid. Node uses the energy of received packets and the available time of links to choose the most reliable route in [6]. Route maintenance will be initiated before the link is to be broken. It is applicable to MANET with preset velocities, whereas in UASN nodes float due to environmental impacts with variable and small velocities.

Moreover, techniques of link interruption prediction and link connection prediction in ubiquitous networks have attracted considerable attention. The literature [7] proposes a probabilistic delay routing algorithm. It uses transmission delay, historical information of link connection to decide different prediction functions for future connection. It focuses on reducing delay but does not consider the impact of environmental changes on link state. In [8], it defines a probability metric-transmission prediction value, when two nodes meet, they exchange respective connection probability list and decide which node will forward data. A method to estimate link lifetime in MANET is given in [9]. It designs a mobile projection trajectory algorithm, whose input is periodically sampled noisy measurements between node pairs on a link. But the prediction precision is decided by the ratio of measurement duration to the whole link lifetime; the greater it is, the more precise the prediction will be. The literature [10] adopts the concept of link availability, it advances an algorithm of estimating continuous link availability between two mobile ad hoc nodes, which uses the distance of nodes to decide the state space of nodes and predict future distance.

The literature [11] is an improvement of [10]. It presents an approach for predicting continuous link availability between two mobile nodes, but it takes radio irregularity and quantities into consideration, better reflecting the reality. Reference [12] advances a constraint-based routing algorithm. It estimates link-state cost before prediction and makes Link-State Advertisements (LSAs) only upon topological changes for reducing communication traffic. The literature [13] proposes a method of Autonomous Underwater Vehicle (AUV) path prediction based on support vector machine, which is effective in the prediction of time sequence characterized by nonlinearity and high dimension. Nevertheless, it is sensitive to the input data and not adaptive to link state changes underwater. Reference [14] presents a model-based monitoring scheme in which each node periodically measures the characteristics of environment

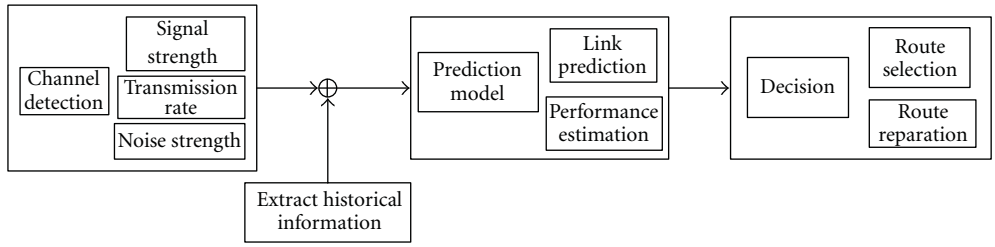


Figure 1: System model of link prediction and route maintenance.

and sends them to gateway. When a link is interrupted, prediction of disconnection duration will be made according to historical information, though it excessively depends on gateway and aggravates the load of gateway.

The literature [15] uses local similarity indices to estimate the likelihood of the existence of links in weighted networks and emphasizes the important role of weak ties. And [16] generalizes the conception of link as the probability of interaction between node pairs in complex networks. It mainly studies how to increase the accuracy of link prediction using a reasonable similarity metric and how to find the potential information of future connection. These literatures handle issues of link prediction in complex networks such as social networks and focus on probability of new links rather than the disconnection of current links, which is not accordant with the application background of our model.

3. System Model

The system model of route selection and maintenance is shown in Figure 1. The key idea is to use channel information, such as signal strength, noise strength, nodes distance, and the historical information of link connection to predict possible link interruption for guiding node's decision.

As UASN is usually sparsely deployed and can be supposed to be relatively static in a certain period, some simplification can be made. So we can consider it as a pseudostatic network, in which link may be intermittently disconnected due to impact of outside objects. This model makes use of channel information by periodic detection and historical information of links to predict route state, then obtains the duration of next connection, and finally makes performance evaluations. Here, for each data transmission, the historical information mainly means the channel state of the last time. In the presence of predicted results, nodes choose an optimal decision go to sleep or select a new route by comparison and judgment. Especially when there is only a single route between two nodes, more historical data is required to get the periodicity of channel changes. Thus, historical information is vital to optimized selection of routes. If there is no historical information or insufficient information, the prediction results we get will not be so precise, because only current channel information is available and further used. Obviously, the more historical information there is, the more precise the calculated periodicity will be. As a result, nodes cannot make precise evaluation of link state until they have obtained enough historical information.

To describe the basic attributes of a link, some definitions are made: D_{ab} is the length of a link, R_{ab} is the data transmission rate (can be taken as a constant here), S_{ab} is the signal strength, N_{ab} is the noise strength, $S_{R\text{-}S}$ is the signal strength threshold of receiver, and E_{ab} is the residual energy.

In the prediction process, some link-state parameters should be defined: S_{ls} is link state indicator, S_{limit} is the threshold of link connection, C_{LK} is the cost of selecting a link, C_{PT_pt} is the cost of selecting a path, P_{con} is the probability of connection, f_{con} is actual connection flag of a link, F_{con} is the predicted connection flag, and L_{con} is the time needed for transmitting a packet.

After making a decision according to prediction results, some estimation parameters are defined: win_l is the predicted duration of link connection, R_b is the reliability of link prediction, P_{er} is the error rate of prediction, and R_{br} is the interruption rate of prediction.

4. Link Prediction and Route Selection Based on Channel Detection

Figure 2 is the flow chart of algorithm implementation on link prediction and route reselection. It mainly includes the following parts: initialization, channel detection, link prediction, and route reparation. This paper assumes that there has been an available routing protocol applied in routing discovery, and the initial route has been established in the stage of initialization. Nodes upload data to base station and detect the acoustic channel periodically.

4.1. Channel Detection. Acoustic channel is a complex and time-varying channel, but there is some information that can be detected and used in link-state estimation, such as signal strength, noise strength, distance between two nodes, and data transmission rate. As for a deterministic network, it is reasonable to regard the data transmission rate as constant in an observation time. Therefore, the information needed for this model can be simplified as three factors: distance of two nodes, signal intensity, and noise intensity.

There are generally three technical ways for measuring distance of node pairs underwater; the first is to use the time difference of signal from the source node to the destination node; the second is to use the round trip time of signal transmission; the third is to use the Received Signal Strength Indicator (RSSI) to estimate the distance between two acoustic transceivers. The first method requires that the nodes have

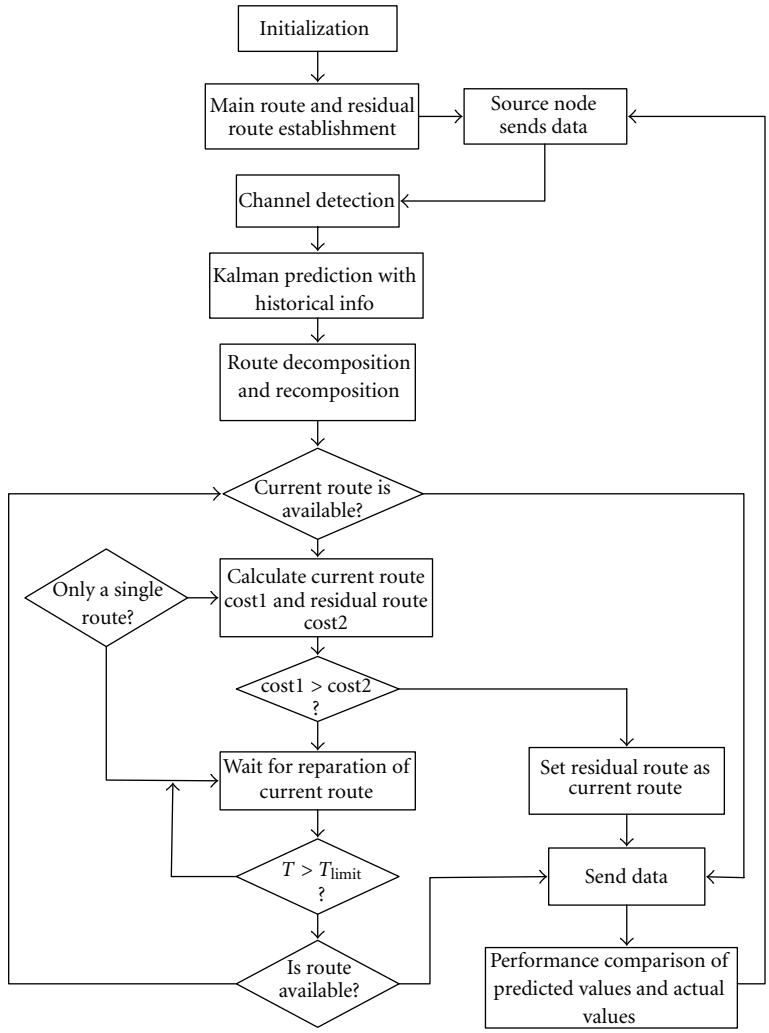


Figure 2: Flow chart of algorithm implementation.

strict clock synchronization, but the transmission delay of acoustic signal is prone to environment, and it is difficult to reclaim and redeploy nodes for clock synchronization. The second means does not require clock synchronization, but the detection packet has to be transmitted twice, which will consume extra node energy and cause long measurement time. Although it is convenient to get the RSSI, but with the effect of multipath and Doppler shift, there is also some uncertainty and inaccuracy in estimating distances with all the three methods. Therefore, a more reasonable method has to be utilized depending on the characteristics of underwater channel. In acoustic channel, propagation loss can be expressed as

$$A(l, f) = A_0 l^k \alpha(f)^l, \tag{1}$$

wherein, l is the distance, f is signal frequency, A_0 is an attenuation constant, k is diffusion factor, $\alpha(f)$ is absorption factor [17]. The relation between receiving power and transmitting power is

$$P_r = A(l, f) \cdot P_t. \tag{2}$$

In the received signal, after identifying and extracting useful information, the rest is noisy signal. By Fourier's transformation or wavelet transformation, the signal can be transformed from time domain to frequency domain, and noise strength can be calculated by signal analysis.

4.2. Link State Prediction

4.2.1. Parameters Definition. (i) Link state indicator. Let S_{ls} be current link state indicator (LSI), which denotes the available signal-noise ratio (SNR).

$$S_{ls} = \frac{S_{ab} - S_{R-S}}{N_{ab}}. \tag{3}$$

(ii) Link state degeneration rate:

$$\beta(k) = \frac{\hat{S}_{ls}(k) - S_{ls}(k-1)}{L_{\text{con}}(k-1)}. \tag{4}$$

$\hat{S}_{ls}(k)$ is the estimated LSI of next cycle by prediction, L_{con} is the data transmission time in *msec*, and $L_{\text{con}} = L_{pk}/R_{ab}$, L_{pk}

is the length of a packet; R_{ab} is data transmission rate which could be regarded as a constant, so data transmission time that depends on L_{pk}. $\beta(k)$ reflects the degenerating rate or enhancing rate of a link in a sampling interval. If $\beta(k) < 0$, it indicates the link state is degenerating and vice versa.

Duration of connection:

$$win\_l = \frac{\hat{S}_{ls}(k) - S_{\text{Limit}}}{|\beta(k)|}, \quad \hat{S}_{ls}(k) - S_{\text{Limit}} > 0, \quad \beta(k) < 0,$$
$$win\_l = \frac{\hat{S}_{ls}(k) - S_{\text{Limit}}}{S_{ls}(k-1) - S_{\text{Limit}}} \cdot L_{con}(k-1), \quad \hat{S}_{ls}(k) - S_{Limit} > 0, \quad \beta(k) > 0, \tag{5}$$
$$win\_l = 0, \quad \hat{S}_{ls}(k) - S_{\text{Limit}} < 0.$$

When link state is degenerating, *win_l* is the time of link state changing from predicted LSI $\hat{S}_{ls}$ to state threshold S_{Limit}, because if S_{ls} is smaller than S_{Limit}, link will be assumed disconnected, and $win\_l = 0$. When link state is enhancing, *win_l* depends on the improvement level of link state and the previous L_{con}. Comparing *win_l* and L_{con}, if $win\_l > L_{\text{con}}$, it indicates current link state can satisfy this data transmission and vice versa.

4.2.2. Prediction of LSI. *Kalman* filter is an optimized autoregressive data process algorithm, which is efficient and suitable for statistical estimation. This model applies the method of *Kalman* filter [18] to predict link state of nodes and gets quite accurate link state as well as relevant link attributes.

$\hat{S}_{ls}(k)$ is the predicted LSI of next cycle by *Kalman* prediction, and then the connection probability of next time is

$$P_{\text{con}}(k) = \min\left(\frac{\hat{S}_{ls}(k)}{S_{ls}(k-1)} \cdot \max[f_{\text{con}}(k-1), \alpha], 1\right). \tag{6}$$

α is a small positive constant to avoid the case that the link of the last cycle was broke (f_{con} is 0), then the successive probability of connection will always be 0.

If $0 \le P_{\text{con}} \le 0.4$, it implies the link will be disconnected next time, so $F_{\text{con}} = 0$. If $0.6 \le P_{con} \le 1$, it denotes the link would be available in the next cycle, so $F_{\text{con}} = 0$. If $0.4 < P_{\text{con}} < 0.6$, it is not sufficient to determine whether link is connected or not in the next cycle, so it will inherit the connection flag of last time, which means $F_{\text{con}}(k) = F_{\text{con}}(k-1)$.

4.2.3. Prediction Result Evaluation. The prediction result evaluation includes two parts: link connection reliability and link state reliability.

(i) Link connection reliability: $L_{crb} = 1 - |P_{\text{con}}(k) - f_{\text{con}}(k-1)|$, and $f_{\text{con}}(k-1)$ is the actual connection flag of time of $k-1$. If it is 1, then link is connected, and if it is 0, then link is disconnected. At time of k, only $f_{\text{con}}(k-1)$ can be known, and we can get the link connection probability $P_{\text{con}}(k)$ of time k. If $P_{\text{con}}(k)$ is very similar to $f_{\text{con}}(k)$, it means the reliability is higher and vice versa.

(ii) Link state reliability

$$Ls_{rb} = \frac{\min\left(\hat{S}_{ls}(k-1), S_{ls}(k-1)\right)}{\max\left(\hat{S}_{ls}(k-1), S_{ls}(k-1)\right)}. \tag{7}$$

LSI is the most important parameter that decides whether a link is connected. Comparing the predicted value with the actual value, the more similar they are, the more precise the prediction is.

The total prediction reliability

$$R_b = \frac{a_1 \cdot Lc_{rb} + a_2 \cdot Ls_{rb}}{a_1 + a_2}. \tag{8}$$

In (8) a_1 and a_2 are positive constants.

(iii) Error rate of link state prediction:

$$S_{er}(k) = \frac{\left|S_{ls}(k) - \hat{S}_{ls}(k)\right|}{S_{ls}(k)}. \tag{9}$$

(iv) Error rate of connection prediction:

$$L_{er}(k) = |P_{\text{con}}(k) - f_{\text{con}}(k)|. \tag{10}$$

The total prediction error rate of prediction:

$$P_{er}(k) = \frac{a_1 \cdot S_{er}(k) + a_2 \cdot L_{er}(k)}{a_1 + a_2}. \tag{11}$$

(v) Interruption rate of a link:

$$R_{br} = \begin{cases} \dfrac{n - \sum_{i=1}^{n} F_{\text{con}}(i)}{n}, & \text{if } 0 < n \le N \\ \dfrac{N - \sum_{i=1}^{N} F_{\text{con}}(i)}{N}, & \text{if } n > N. \end{cases} \tag{12}$$

The maximal number of historical data recorded in each node is N. If running times are smaller than N, interruption rate of a link is calculated by data from the initial time to current time. If running times are more than N, interruption rate is calculated by data of the last N times.

(vi) Energy loss estimation:

It is assumed that input parameters are the same both with and without link prediction. When the main route is confirmed to be broken, a new route must be reselected. Since no prediction result can be used, the newly selected route may still include some broken links, which also cannot guarantee reliable data transmission. And during the time of waiting for confirmation T_{cfm}, no matter whether link is broken, the source node will send inquiring packets by period of T_t; the energy needed of each transmission is e_d, so the energy consumed while waiting is $T_{cfm} \cdot e_d / T_t$. Reselecting a route consumes e_{rt}. If there is a single available route when links are disconnected, nodes have to wait for link repairing.

If there is link prediction, nodes will know the approximate duration of link interruption, and they can choose to go to sleep mode while waiting for link reparation or immediately reselect an available link to substitute the old

one. So we define that the power consumption in sleep mode is P_{slp}, the predicted duration of link interruption is T_w, and the energy needed for a cycle of sleep-mode is $T_w \cdot P_d$.

There are 4 cases concerning single route and multiple routes with or without prediction:

(1) single route without prediction

$$E_{g1} = \left(1 - 0.5^{lk\_num}\right) \cdot \left(P_{cfm} \cdot T_{cfm} + e_d \cdot \frac{T_w}{T_t}\right), \quad (13)$$

(2) single route with prediction

$$E_{g2} = (1 - P_{er}) \cdot P_{slp} \cdot T_w + P_{er} \cdot E_{g1}, \quad (14)$$

(3) multiple routes without prediction

$$E_{g1} = P_{cfm} \cdot T_{cfm} + e\_rt + P_{br1} \cdot \left(P_{cfm} \cdot T_{cfm} + e\_rt\right), \quad (15)$$

and $P_{br1} = (1 - 0.5^{lk\_num})^{rt\_num-1} + 0.5^{lk\_num} \cdot (1 - 0.5^{lk\_num})$ is the probability of new route still including broken links.

(4) multiple routes with prediction

$$E_{g2} = e_{rt} + P_{br2} \cdot (e_{rt}), \quad (16)$$

and $P_{br2} = (1 - (1 - R_{br})^{lk\_num})^{rt\_num-1}$ is the probability of the new route still including broken links.

4.2.4. Link Prediction for Single Route. The practical application scenes probably include the following aspects. First, in the ocean, communication interruption may be caused by ships passing by, yet the time of ships passing by is not occasional but regular to some extent. Second, the tide also has impact on acoustic communication, and it has clear periodicity. Last but not the least, the probability of fish flocks that traverse an area complies with the movement regularity of fish.

If there is only one route between the source and destination with broken links, the source node has to wait for link reconnection. At this time, the model will predict the disconnection duration from current time to the next connection, which is the time that node needs to wait for. In order to make reasonable prediction and illuminated from the fact that periodic phenomena aforementioned, here we assume that link interruption is a kind of periodical events. After the first period, there will be records of link interruptions and other channel information. Then node can estimate the duration of link interruption based on the records of the last period. Therefore, the source node can resend data after a period of sleeping. This method not only reduces the energy consumption of nodes for continuous attempt of sending requiring packets but also guarantees the effectiveness and security of data transmission.

Periodic signal analysis can be carried out by differentiation analysis, polynomial fitting, Fourier's analysis, and wavelet analysis [19]. Because LSI in this model is a linear composition of each link attribute without complex trend component, Fourier transformation can directly get the frequency of signal.

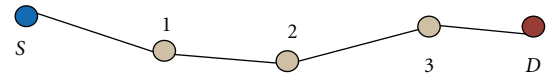


Figure 3: Data transmission of a single route.

Long-time Fourier transform requires the entire or a large part of time-domain signal to obtain the frequency characteristics. Consequently, this needs large storage space to save the time-domain signal. However, prediction using fewer signals may bring in large prediction error. In this case, wavelet transform or short-time Fourier transform [20] possesses inhere advantages. So short-time Fourier analysis is adopted here for nodes to find periodicity of channel changes.

4.3. Route Selection. When the source node decides to initiate the route reselection, it will have to evaluate the cost of a path. But the source node must get the state information of all nodes along the path to decide which back-up route will become the new main route. Here, we advance a concept of route decomposing and recomposing to simplify data delivery process and then choose a new route with the least cost.

4.3.1. Route Decomposing and Recomposing Hop by Hop. As Figure 3 shows, data from source S reaches sink D by 3 hops, S can communicate with node 1 and acquire its data, node 1 can acquire node 2's data, and node 2 can acquire node 3's data. Indirectly, S gets D's data, which contains the state information of link $3 \rightarrow D$. Similarly, S can get all the other links state on the path. To decrease the communication load of S, this model innovatively decomposes a route hop by hop and then recomposes them in a reverse sequence. The detailed process is, from the sink node D in the reverse direction, in light of node 2, node 3 and D can be seen as a new node 3+; in light of node1, node 2, and 3+ can be seen as a new node 2+; similarly, node 1 and 2+ can be seen as a new node 1+. Therefore, node S only needs to communicate with the new node 1+. In this way, the process of communication in a route is simplified.

If S cannot communicate with 1 at a time, so the data of 1+ about the link state from $1 \rightarrow D$ cannot be relayed to S, and S cannot estimate the state of the whole route. At this time, node 1 must search for other routes to S, For example, 1 finds 4 to forward data to S, and this process is called local rerouting, as Figure 4 shows.

This model divides links into hot links and common links as shown in Figure 4. The link that is included in more than 2 routes is seen as a hot link; its interruption will lead to many routes failure. Link $3 \rightarrow D$ in Figure 4 is a hot link. If $3 \rightarrow D$ is broken, source node S has to wait for route repairing to restore normal transmission. Therefore, it is necessary to predict the duration of hot link interruption to avoid node's repeated attempt of connection.

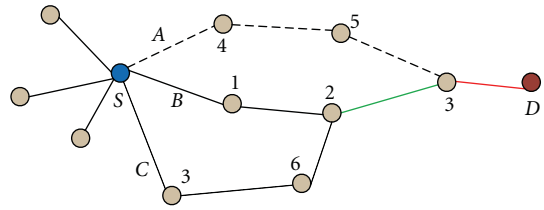


Figure 4: Demonstration of hot links.

4.3.2. Routes Reselection. When a route was interrupted, the source node selects an available and stable route from the main route and all back-up routes for data transmission. Here, we integrate all factors that influence link reliability and put them in the form of cost. Route with the smallest cost will be chosen as the main route.

Assuming that the main route is B, and B is broken at this moment (as shown in Figure 4), so it demands the steps of link prediction and route reselection. Source node will select a route with the smallest cost from the back-up routes A, C, and B for the next transmission.

The cost of a link is

$$C_{LK} = k1 \cdot \frac{D_{ab}}{R_c} + k2 \cdot \frac{\text{hop\_num}}{\text{hn\_num}}, \tag{17}$$

where R_c is the communication range of a node, hn_num is the maximal hop number on a route, and hop_num is the number of hops between the link and the destination node (D) along a route. $k1$ and $k2$ are positive constants bolw 1 and $k1 + k2 = 1$. The cost of a path is

$$C_{PT} = C_{LK}(1) \cdot C_{LK}(2+) = C_{LK}(1) \cdot C_{LK}(2) \cdot C_{LK}(3+) = \cdots = \prod_i C_{LK}(i), \tag{18}$$

i is the ith link along the route. As the connectivity of a route depends on the connectivity state of links along it, so $C_{PT} = C_{PT} - \min(\widehat{S}_{ls})$, $\widehat{S}_{ls}$ is the LSI set of all links on a route. If there are broken links on the route, C_{PT} must be multiplied by a factor α according to the number of broken links to increase the cost, which is $C_{PT} = C_{PT} \cdot \alpha$, α is a constant greater than 1.

This can ensure that a route with more broken links will produce a higher cost. As a result, it is less impossible for a path with broken links to become the new route. If all routes have some broken links, the prediction of disconnection duration will be needed, as Section 4.2 describes.

4.3.3. Back-Up Routes Maintenance. If the main route is connected and available all the time, there will be no chance for back-up routes to transmit data, so they could not acquire the historical information of links along the route and channel characteristics. When one back-up route is selected as a new route, since it does not have any historical information, it cannot make accurate prediction of link states. Therefore, in order to get necessary historical information for back-up routes, we adopt a fresh idea of state maintenance on back-up routes. Two strategies are put forwarded. (i) Back-up routes are intermittently used to transmit data. (ii) Nodes along back-up routes are intermittently used to detect the acoustic link.

In detail, there are two techniques of intermittently using back-up routes. First, sensing data is transmitted by back-up routes once in a while. Second, less important data or data with lower delay constraint is transmitted by back-up routes. The first method is mainly applied in situations where data has similar priority or importance, and the second is suitable in situations where data packets have different priorities, especially when there are emergent event messages. At this time, their priority is higher than common data packets, so they should be transmitted via the main route, and other data packets are transmitted via back-up routes.

Also, there are two means of intermittently detecting the acoustic channel by back-up routes. First, when this back-up route is used to transmit data, nodes along the route simultaneously detect channel. Second, no matter whether the route is used, nodes along the route always periodically detect channel.

5. Simulation and Analysis

To verify the prediction model and route selection algorithm in case of link interruptions and route interruptions, two different situations are classified: prediction for multiple routes and prediction for a single route. Here we use Opnet and Matlab as the simulation platform.

Simulation setup. The channel model is flat Rayleigh fading, carrier frequency is 20 kHz, the modulation is Quadracture Amplitude Modulation (QAM), data transmission rate is set to 1000 bits/s, and the length of packet is changed from 100–400 bits. Sensor nodes gather data and transmit data to sink node in the scheme of event-driven, which means that if nodes' data satisfies the need of one transmission or nodes receive the transmission command, it will begin to find a route and transmit data.

5.1. Prediction for Multiple Routes. For simplicity, we assume a small scale network with 12 sensor nodes. There are three available routes between source node and destination node after routing discovery, and the initial main route is the red line as Figure 5 shows. The input parameters are the attributes of acoustic channel; by formula (3), they are transformed to a parameter LSI, and the actual LSI S_{ls} is demonstrated in Figure 6(a).

Figures 6 and 7 are predicted LSI, connection flag, error rate of prediction, reliability, interruption rate, and connection duration on a link. Figure 6(a) is the comparison of actual LSI and predicted LSI by *Kalman* filter. Figure 6(b) is the actual flag of link connection f_{con} and predicted flag F_{con}, it is evident to see that estimation results match with the real values. P_{er} in Figure 6(b) is the error between F_{con} and f_{con}, when LSI is near the threshold S_{limit}, the predicted LSI will deviate from the actual value, because at this point connection flag f_{con} suddenly changes from 1 to 0 or 0 to 1.

Figure 7(a) is link interruption rate and prediction reliability; the former value is floating from 0.8 to 1.0; the latter one is varied with the changes of F_{con}, while t = 35–50,

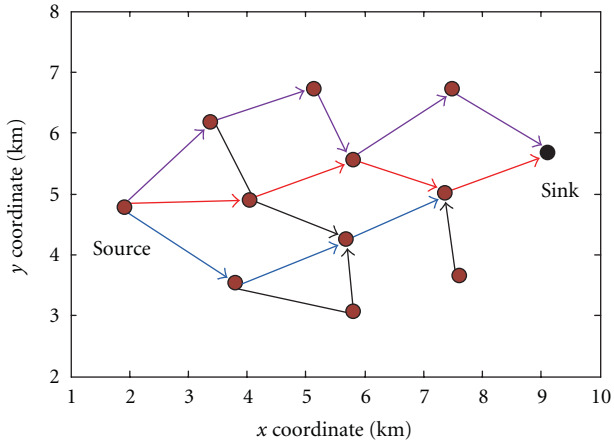


Figure 5: Initial topology.

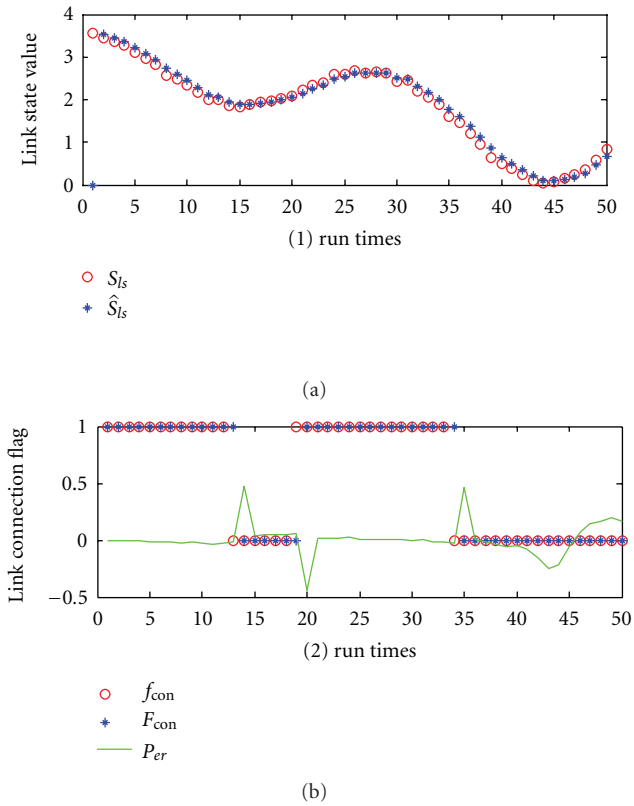


Figure 6: LSI prediction and link flag.

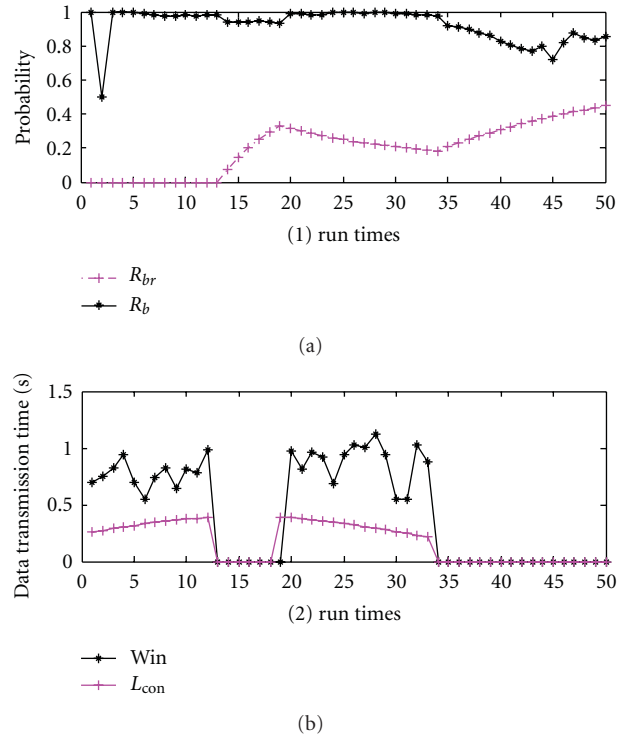


Figure 7: Comparison of predicted parameters.

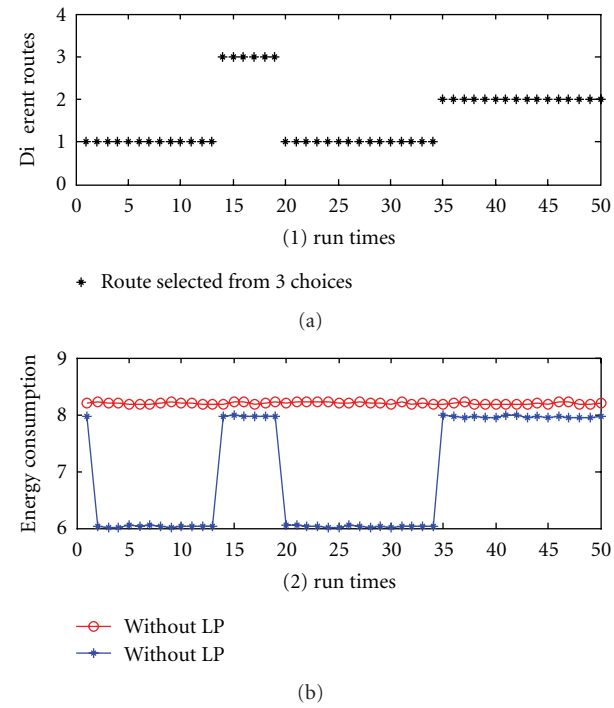


Figure 8: Route selection and energy consumption.

the link is continuously broken, R_{br} is keeping increasing, but once link reconnects, it will keep decreasing. Figure 7(b) is the comparison of predicted duration of link connection and the actual value; while LSI is high, the predicted win_l is relatively large, but when LSI is below S_{Limit}, win_l is sharply decreased to 0. Besides, for L_{con} is the actual time required for each data transmission, as long as $win\_l > L_{\text{con}}$, it is thought that this transmission can be completed.

Figure 8(a) is the different main routes selected by nodes after running 50 times. The numbers on the vertical axis represent three different paths. It is obvious that route 1 is most frequently used, and route 3 is scarcely used. While $t = 35$–50, because route 1 has a broken link and route 3 is not so efficient, route 2 is chosen.

Figure 8(b) is the comparison of energy consumed with and without link prediction. It can be seen that the energy

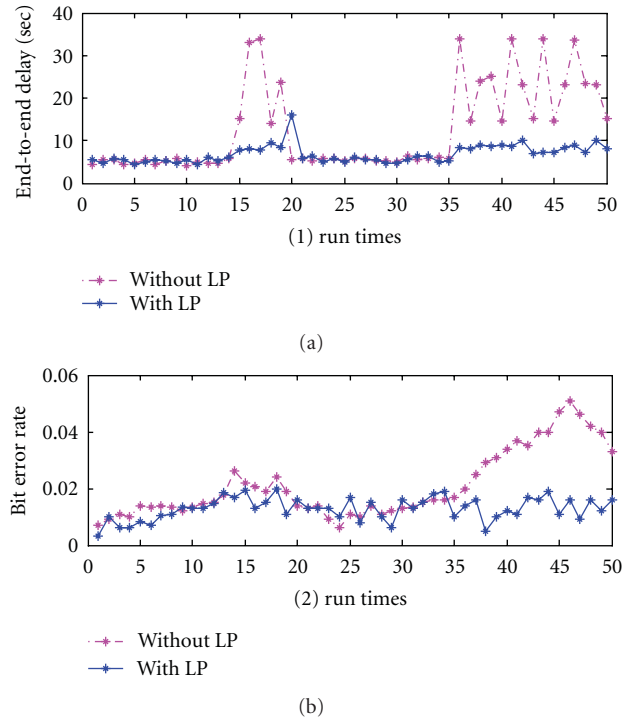


Figure 9: End-to-end delay and bit error rate.

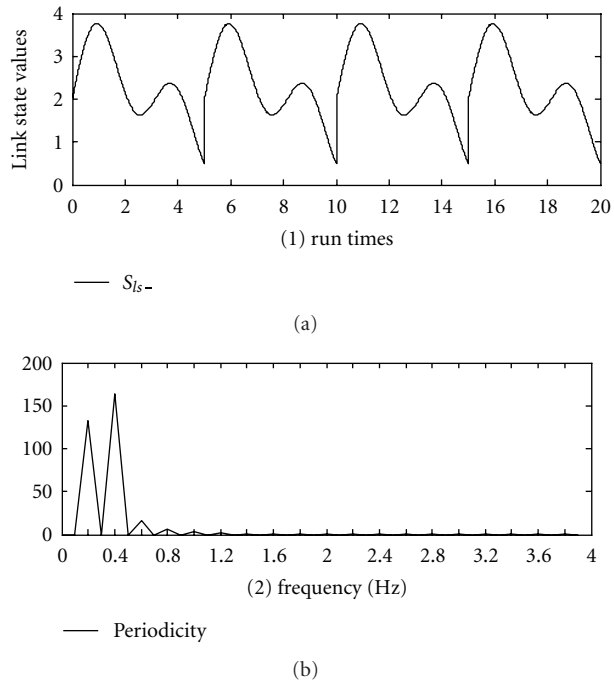


Figure 10: Periodicity detection.

consumed without prediction is more than 30% higher than that with prediction. This is because when nodes estimate that the link will be disconnected, they will reselect a most potential route without broken links or with fewest broken links. And the better the link state is, the less possible broken links are to be included in this route. As a result, the energy consumption of node is significantly decreased. However, without link prediction, nodes have to select a new main route repeatedly if it still contains several broken links. This will consume much more energy and may lead to another data transmission failure.

Figure 9(a) compares the end-to-end delay with link prediction (LP) and without link prediction. With reference to Figure 6, when link state is good and there is no interruption, delay in the two cases is almost the same. And when there is link disconnection, delay with LP is a little longer than that when there is no link disconnection. This is because nodes consume some processing time when reselecting routes. But without LP, nodes can affirm link interruption after the threshold of waiting time – T_w, which consumes a lot of time. Besides, it is possible that the reselected route is not the optimal, which means that the new route may also include disconnected links. So the end-to-end delay is much longer than that with LP. The impulse in the figure is caused because of the prediction error when link starts to be broken and reconnects. This costs some waiting time of nodes.

From the red line in Figure 9(b), we can see that, with the changes of S_{ls}, the bit error rate (BER) is changing. When S_{ls} is decreasing, BER gradually increases and vice versa. When t = 0–10, link state is almost the best, so BER is almost decreased to be the lowest and when 4–4.5, BER is almost increased to the greatest. Therefore, the BER we derive is accordant with the actual link state. This is the result without LP.

When there is LP, as nodes can predict link disconnection, they will choose another path to transmit data. For link interruption that is because environmental changes cause acoustic changes and make the data error rate in the receiver larger than the tolerant values (such as BER $\geq$ 0.02), so it is assumed that link has been broken and not suitable to transmit data. Accordingly, the receiver will obtain data with high BER in this case. But the new route reselected contains less broken links as much as possible, hence by the BER is greatly decreased than that without LP.

From the aspects of energy consumption, end-to-end delay, and bit error rate, it is obvious that the mechanism proposed in this paper is effective and efficient.

5.2. Prediction for Single Route. If there is only one route between the source and destination nodes but with disconnected links, source node has no choice but to wait for link repairing. Yet we can estimate the duration of route interruption in order to let nodes go to sleep for energy conservation. In Section 4.2.4, we have demonstrated that real environmental changes are often quasiperiodic events. Thereby, it is reasonable to suppose that actual link state has inherent periodicity. When a link interrupts, it does not know the current time but only can estimate the periodicity of link changes by historical records of LSI and find its relative position in a period.

Figure 10(a) is the simulated changes of link state, in which there are 4 periods and the periodicity is 5. Figure 10(b) is signal frequency derived from Fourier

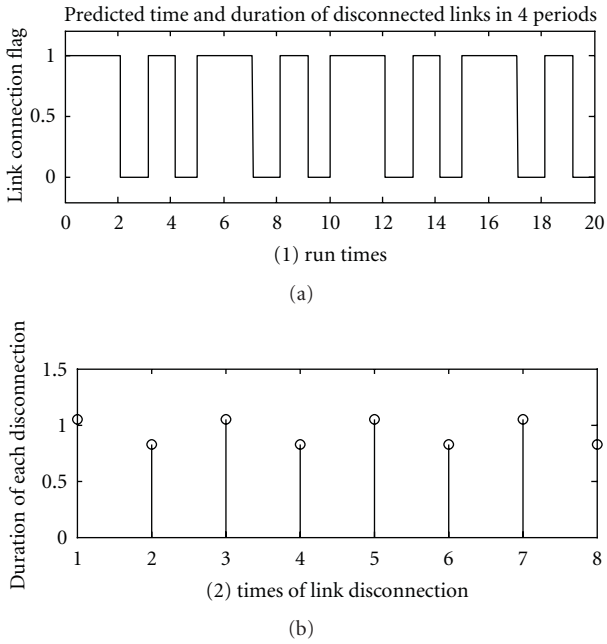


Figure 11: Periodic interruption of a link.

transformation, it can be seen that there is a pulse in $f = 0.2$ Hz, which denotes that the basic frequency of signal is $f = 0.2$ Hz and a multiple frequency is $f = 0.4$ Hz. So we can deduce the signal periodicity is $T = 1/f = 5$. Figure 11(a) is the relation between point of a link beginning to interrupt and the duration of interruption in each period, if the flag of link connection is 1, then the link is connected and vice versa. In Figure 11(b), the horizontal axis is the number of times of link interruption; the vertical axis is the length of each interruption, corresponding to the Figure 11(a).

The two figures indicate that it is convenient to detect the periodicity of link state by simple signal transformation. And this method can improve the accuracy and efficiency in the process of link prediction.

6. Conclusion and Prospect

In this paper, we have presented a link prediction and route reselection model of DTN in UASN to facilitate the data transmission in case of link interruption. The primary contributions of this paper are summarized as follows.

(1) The characteristics of acoustic channel are taken into account in UASN, and channel state is periodically detected to derive the LSI. Depending on historical link state information and current channel information, the LSI of next time can be predicted to determine whether link will be interrupted.

(2) A method of route decomposing and recomposing hop by hop is proposed for the optimization of route selection. For the effectiveness of back-up routes, we advance a method of back-up routes maintenance to keep them with fresh channel information.

(3) With respect to multiple routes and single route, we propose a mechanism of route reselection and route maintenance. Specifically, in the case of single route, we advance the idea to utilize the periodicity of environmental changes to help predict link interruption.

The simulations compare three network performances between two cases with and without link prediction. The results prove that the method of link prediction and route reselection is valid and effective by analysis on the energy consumption of nodes, end-to-end delay of data transmission and the bit error rate of the receiver.

Although link prediction has become a research hot spot in cognitive radio, complex networks, there are still not deep explorations on the relation between algorithm performance and network structure. Moreover, no uniform standards are defined on the link characteristics. Hence, link prediction, especially in UASNs, needs continuous research effort. This paper applies link prediction into underwater pseudostatic sensor network to improve the data transmission in challenged environment, which is a typical application scene. How to further optimize and generalize the model and expand the application scope is a research direction of future work. Moreover, how to integrate link prediction into the process of route discovery is a potential aspect of next stage.

Acknowledgments

The authors would like to thank the reviewers for their detailed comments that have helped to improve the quality of the paper. This work was funded in part by National Basic Research Program of China (973 Program) 2011CB707106, NSF China 60972044 and the Fundamental Research Funds for the Central Universities 201121202020002.

References

[1] F. Khadar and D. Simplot-Ryl, "From theory to practice: topology control in wireless sensor networks," in *th 10th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '09)*, pp. 347–348, May 2009.

[2] Y. Li, X.-Y. Wang, Z.-J. Liu, and Y.-H. Zhou, "Analysis of link interruption characteristics in the DTN," *Journal on Communications*, vol. 29, no. 11, pp. 232–236, 2008.

[3] Z. Guo, G. Colombi, B. Wang, J. H. Cui, D. Maggiorinit, and G. P. Rossi, "Adaptive routing in underwater delay/disruption tolerant sensor networks," in *the 5th Annual Conference on Wireless on Demand Network Systems and Services (WONS '08)*, pp. 31–39, January 2008.

[4] H. Samuel, W. Zhuang, and B. Preiss, "DTN based dominating set routing for MANET in heterogeneous Wireless Networking," *Mobile Networks and Applications*, vol. 14, no. 2, pp. 154–164, 2009.

[5] S. Agarwal, A. Ahuja, J. P. Singh, and R. Shorey, "Route-lifetime assessment based routing (RABR) protocol for mobile Ad-hoc networks," in *IEEE International Conference on Communications*, pp. 1697–1701, June 2000.

[6] L. Hong, T. P. Huang, W. X. Zou, S. R. Li, and Z. Zhou, "Research of AODV routing protocol based on link availability prediction," *Tongxin Xuebao/Journal on Communications*, vol. 29, no. 7, pp. 118–123, 2008.

[7] L. Yin, H. M. Lu, and Y. D. Cao, "Probabilistic delay routing for delay tolerant networks," in *the 10th International Conference on Advanced Communication Technology*, pp. 191–195, February 2008.

[8] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," in *Proceedings of the 1st International Workshop Service Assurance with Partial and Intermittent Resources (SAPIR '04)*, vol. 3126 of *Lecture Notes in Computer Science*, pp. 239–254, Fortaleza, Brazil, August 2004.

[9] Z. J. Haas and E. Y. Hua, "Residual link lifetime prediction with limited information input in mobile ad hoc networks," in *the 27th IEEE Communications Society Conference on Computer Communications (INFOCOM '08)*, pp. 26–30, April 2008.

[10] L. Gong, Y. Bai, M. Chen, and D. Qian, "Link availability prediction in Ad Hoc Networks," in *the 14th IEEE International Conference on Parallel and Distributed Systems (ICPADS '08)*, pp. 423–428, December 2008.

[11] Y. Bai and L. Gong, "Link availability prediction with radio irregularity coverage for mobile multi-hop networks," *IEEE Communications Letters*, vol. 14, no. 6, Article ID 5474929, pp. 518–520, 2010.

[12] X. Masip-Bruin, E. Marín-Tordera, M. Yannuzzi, R. Serral-Gracià, and S. Sànchez-López, "Reducing the effects of routing inaccuracy by means of prediction and an innovative link-state cost," *IEEE Communications Letters*, vol. 14, no. 5, Article ID 5456076, pp. 492–494, 2010.

[13] R. Zhang, G. Liu, and X. Li, "Real time path predicting for autonomous underwater vehicle using support vector regression machines," in *the 4th International Conference on Natural Computation (ICNC '08)*, pp. 414–416, October 2008.

[14] D. Tulone, "A model-based monitoring scheme for disruption-tolerant underwater sensor networks," in *the 6th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, June 2009.

[15] L. Lü and T. Zhou, "Role of weak ties in link prediction of complex networks," in *the 1st ACM International Workshop on Complex Networks in Information and Knowledge Management (CNIKM '09)*, pp. 55–58, November 2009.

[16] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[17] F. Niu, Y. Yang, and C. Guo, "Simulation study on time-varing multiple paths of acoustic channel," *Jounal of Oceanography in Taiwan Straint*, vol. 28, no. 4, pp. 586–591, 2009.

[18] J. F. Xue, H. M. Lu, and L. Shi, "Design of DTN routing algorithm based on probabilistic delay," *Transaction of Beijing Institute of Technology*, vol. 28, no. 8, pp. 687–691, 2008.

[19] H. Wen and Z. Zhang, "Detection of the periodic signal in the deformation analysis with wavelet and fourier tansformation," *Bulletin of Surveying and Mapping*, no. 4, pp. 14–16, 2004.

[20] R. Zhao and P. Xiong, "Seperation of periodic signals using variable detection step," *Transactions of Sichuan University(Natural Science Edition)*, vol. 32, no. 5, pp. 527–532, 1995.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 126407, 12 pages
doi:10.1155/2011/126407

Research Article

SEF: A Secure, Efficient, and Flexible Range Query Scheme in Two-Tiered Sensor Networks

Jiajun Bu,[1] Mingjian Yin,[1] Daojing He,[1] Feng Xia,[2] and Chun Chen[1]

[1] *Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310007, China*
[2] *School of Software, Dalian University of Technology, Dalian 116024, China*

Correspondence should be addressed to Jiajun Bu, bjj@zju.edu.cn

Received 10 February 2011; Accepted 19 May 2011

Large-scale wireless sensor networks follow the two-tiered architecture, where master nodes take charge of storing data and processing queries. However, if a master node is compromised, the information stored in it may be exposed, and query results can be juggled. This paper presents a novel scheme called SEF for secure range queries. To preserve privacy, SEF employs the order-preserving symmetric encryption which not only supports efficient range queries, but also maintains a strong security standard. To preserve authenticity and integrity of query results, we propose a novel data structure called Authenticity & Integrity tree. Moreover, SEF is flexible since it allows users to include or exclude the authenticity and integrity guarantee. To the best of our knowledge, this paper is the first to use the characteristic of NAND flash to achieve high storage utilization and query processing efficiency. The efficiency of the proposed scheme is demonstrated by experiments on real sensor platforms.

1. Introduction

The traditional architecture of wireless sensor networks (WSNs) always assumes that a trusted base station (or sink) is present and responsible for collecting data from the sensor nodes and processing query requests from the users. However, many WSNs are deployed in hostile and harsh environments such as battlefields, forests, and oceans where it is impossible or difficult to establish a stable communication link from sensor nodes to the base station. Summarizing the above, we can see that such an architecture only makes sense for experimental and small networks. As illustrated in Figure 1, large-scale WSNs follow a two-tiered architecture, where a large number of regular resource-poor sensor nodes are at the lower tier and fewer resource-rich master nodes are at the upper tier. The two-tiered architecture has been widely adopted [1, 2] because of the benefits of energy and storage saving as well as the efficiency of query processing. Compared with sensor nodes, master nodes have more powerful computing power and more energy supply, additionally they are equipped with several gigabytes of NAND flash for tens of dollars. The sensor nodes are organized as cells and each cell includes a master node referred to as the cell header. The sensor nodes are responsible for sensing data and submitting data to the master nodes, while the master nodes take charge of storing data and performing resource-demanding tasks such as query processing. In this paper, we focus on range queries, namely, once the base station launches a query request to a master node to query the data items in the range [low, high], the master node must return all data items in the query range. Note that if *low* is equal to *high*, the query is actually an equality query that requires the master node to return all data items that are equal to *low* (or *high*). Our scheme also supports equality queries. For simplicity, we refer to the sensor node, master node, and base station as SN, MN, and BS, respectively, in the rest of this paper.

The inclusion of MNs also incurs significant security challenges. Since both data storage and query processing rely on MNs, MNs are prone to be the target of attacks, especially in some unattended and hostile environments such as military scenarios. And the adversary is more attracted to compromise MNs to greatly damage WSNs. If MNs are compromised, the adversary can launch at least three types of attacks. First, the adversary can read all data stored in the compromised MNs and breach data privacy. A sound defense to this attack should ensure data privacy and still enable efficient query processing. Second, the compromised MNs may send tampered or forged data in response to queries,

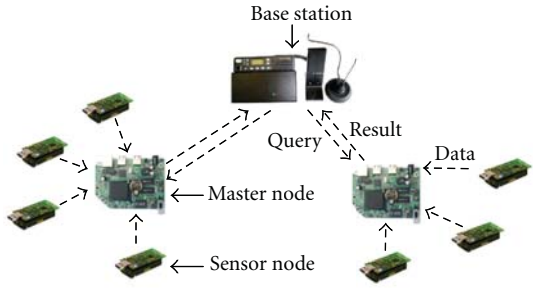


Figure 1: Architecture of a two-tiered WSN.

which breaches the authenticity of query results. Thirdly, the returned query results may not be integral, which means that some qualified data items are omitted by the compromised MNs. Therefore, a solution is required urgently, which can offer efficient and effective range queries. At the same time it can also preserve data privacy as well as the authenticity and integrity of query results at the presence of compromised MNs.

In some cases, the authenticity and integrity guarantee might not be always necessary, for example, out of consideration for reducing communication and verification cost. Therefore, a range query scheme should be flexible, namely, the users can choose to include or exclude the authenticity and integrity guarantee according to different applications.

The main contributions of this paper include the following: (1) To the best of our knowledge, this paper, for the first time in the literature, considers the characteristic of NAND flash when designing query schemes in WSNs. Our work explores the storage media concerns in general WSNs and provides a novel approach for data storage and query processing. (2) We propose a secure, efficient, and flexible range query scheme called SEF for two-tiered WSNs, in which only the order information of the encrypted data is exposed to MNs. (3) Since some applications may require authenticity and integrity guarantee of query results, and others may not, by keeping encrypted data separated from verification information, our scheme provides much flexibility. (4) We evaluate our scheme in a test bed; the results show that compared with existing schemes, our scheme performs better on both energy consumption and storage consumption.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related work. Section 3 presents the network and adversary models. Section 4 introduces some techniques and notations used in this paper. Section 5 describes our scheme in detail. Section 6 gives a detailed security analysis of the proposed scheme. Section 7 reports the performance evaluation results. Section 8 concludes the paper.

2. Related Work

Data security has gained some attention in outsourced database community [3–5]. In an outsourced database system, a data owner publishes his/her data through a remote server which may be untrusted or compromised. The remote server answers user queries on behalf of the data owner. However, due to the unique and challenging characteristics and security requirements of WSNs, such as limited computing power and communication capability, these approaches are not suitable for WSN applications.

Despite significant progress in WSN security such as [6, 7], secure range queries have started receiving people's attention very recently [8–12]. In [8] Sheng and Li first considered the privacy issue in the design of range query schemes in WSNs. We call it "SL" scheme in the rest of this paper. They apply the bucketing technique introduced in [13] to fulfill secure range queries. The basic idea of bucketing technique is to partition the domain of sensed data values into buckets and each bucket is assigned with a bucket tag. The bucket tags are exposed to SNs, MNs, and the BS, but how to partition the domain of sensed data values into buckets is only known to SNs and the BS. A data item sensed by a SN is placed into a bucket according to the partition. Then all data items falling into the same bucket are encrypted as a whole, at last the SN sends the buckets as well as the bucket tags to its MN. In order to make the BS verify whether the returned query result is integral, [8] introduces the encoding technique. The bucket into which no data item falls is attached with an encoding number.

In [9, 10], Shi et al. proposed an optimized integrity verification version of SL scheme, to reduce the communication cost between the SN and its MN caused by encoding technique. The basic idea is that each SN uses a bit map to represent which buckets have data and broadcasts its bit map to nearby SNs. Each SN attaches the bit maps received from other SNs to its own data items, then encrypts them together. However, the optimization technique causes a serious problem that is a compromised SN can easily breach the integrity verification of the network by sending false bit maps. On the contrary, SL scheme and our scheme do not have the problem. As the techniques used in [9, 10] are similar to [8] except the optimization for integrity verification, they inherit the same weakness with SL scheme.

There are many common drawbacks in the schemes of [8–10]. (1) They require in advance the accurate distribution of sensed data items, $F(x)$ (i.e., the probability that a sensed data item is x). In reality, it is very hard to meet the requirement; sometimes the BS is unable to estimate $F(x)$ in advance. (2) The bucketing technique incurs a problem of false positive [13]; some useless data items are sent back to the BS. The less the buckets are, the more useless data items are. While the more the buckets are, the more the exposed privacy is. Consider an extreme case, every distinct value has a unique bucket tag. If a SN is compromised, the value and bucket-tag mapping will be exposed, then the adversary can derive all data values stored in the compromised MN, even though the data is encrypted. (3) As all SNs in the same cell share the same bucket tags, and the bucket tags are constant, if the adversary compromises a SN, the adversary can get other SNs' information just by overhearing. Specifically the information that into which range the data items fall, as well as how many data items a range has, is exposed to the adversary. (4) They do not consider the storage media issue of MNs. Since SNs or MNs can create a search index, such

as B+ tree [14] and EMB tree [3], to speed up the process of searching data, the cost of reading data from NAND flash is dominant in query processing. The I/O interface of NAND flash does not provide a random-access external address bus, data must be read on a pagewise basis, with typical page sizes of 256 or 512 bytes. The storage media issue greatly influences the storage utilization and query processing efficiency.

In order to address the drawbacks of these schemes, Chen and Liu proposed a novel privacy and integrity preserving rang query protocol called SafeQ-Basic in [11, 12]. To preserve privacy, their protocol uses the prefix membership verification scheme first introduced in [15]. The basic idea of the prefix membership verification scheme is to convert the verification of whether a number is in a range to several verifications of whether two numbers are equal. Interested readers can refer to [15] for more detailed information. To preserve integrity, they propose a data structure called neighborhood chaining. The main idea of neighborhood chaining is to sort the sensed data items in ascending order first, and each data item d_i concatenates its right neighbor d_{i+1}, then $d_i \| d_{i+1}$ are encrypted together, $i = 1, 2 \ldots, \mathrm{N} - 1$ (suppose there are N data items), where $\|$ denotes concatenation.

SafeQ-Basic causes large computation, communication, and storage overhead, because the prefix membership verification scheme requires many HMAC operations and produces a lot of HMACs. To reduce the communication and storage overhead, [11, 12] also proposed an optimized version of SafeQ-Basic, called SafeQ-Bloom. SafeQ-Bloom uses a Bloom filter to represent HAMCs. Thus, an SN only needs to send the Bloom filter instead of HMACs. The number of bits needed to represent the Bloom filter is much smaller than that needed to represent HMACs. However, to produce the Bloom filter further leads to more computation overhead. These features make SafeQ-Basic and SafeQ-Bloom not efficient and not suitable for WSNs. We find that though SafeQ-Basic and SafeQ-Bloom apply the ordinary symmetric encryption algorithm to encrypt sensed data items, such as DES in their experiments, the order of encrypted data items is still exposed to MNs because all encrypted data items are sorted in ascending order. This security standard is the same as that of SEF. Moreover, neither SafeQ-Basic nor SafeQ-Bloom considers the storage media issue.

3. Network and Adversary Models

3.1. Network Model. We consider a large-scale two-tiered WSN as illustrated in Figure 1. The WSN consists of SNs and MNs; the network is partitioned into cells, according to the geographic location, each SN belongs to one cell, and each cell contains an MN which is in charge of other SNs in the cell. In some scenarios, two adjacent cells maybe overlap; in this case the SNs in the overlapping regions are affiliated with both MNs. The SNs are limited in storage, energy supply, and computing power; in contrast the MNs are resource-rich devices.

Time is assumed to be divided into time slots and all SNs and MNs are loosely synchronized with the BS. We assume that every SN senses the environment data in a fixed rate and periodically submits sensed data to its MN.

Contrary to conventional WSNs, our WSN has no stable always-on communication link to the external BS. MNs store all data collected from SNs which they are in charge of. The BS can query MNs for data on demand. We assume that MNs have enough storage to store data. Nowadays, most embedded devices use NAND flash as the storage media. In this paper, we follow the assumption that the storage media of MNs is the most conventional NAND flash [16, 17].

In this paper, for sake of simplicity, we assume that any query request can be converted into multiple range query primitives having the format:

$$\text{cell} = C \wedge \text{SN} = S \wedge \text{time} = T \wedge \text{value} \in [\text{low}, \text{high}], \quad (1)$$

where C, S, and T represent the IDs of a cell, SN, and time slot, respectively, and [low, high] is the query range.

3.2. Adversary Model. Due to the broadcast nature of wireless communication environments, the adversary may launch many attacks such as jamming, passive eavesdropping, and bogus-message injection. There exist rich elegant defenses to these attacks such as in [18–23]. We resort to the existing rich literature for these attacks. This paper particularly focuses on SNs compromise and MNs compromise attacks.

(1) SNs compromise: if an SN is compromised, the data values stored in the SN will be exposed to the adversary and the compromised SN may send forged data to its MN. Unfortunately, it is very difficult to prevent such attack without the use of tamper proof hardware. We follow the same SNs compromise assumption with [11, 12]. If an SN is compromised, the adversary can just access the data stored in the compromised SN. It must be guaranteed that the adversary cannot make any damage to other SNs and MNs by the compromised SN.

(2) MNs compromise: once an MN is compromised, the adversary can read all data stored in the compromised MN and try to obtain the data values, which violates data privacy. Second, the compromised MN can return tampered, forged or not integral query results in response to queries. As compromising an MN can cause greater damage to WSNs than compromising an SN, the adversary is more attracted to compromise MNs. We propose a novel data structure called Authenticity&Integrity tree (AI tree for brevity) to guarantee the authenticity and integrity of query results from MNs.

4. Preliminaries and Notations

In this section we review some cryptographic essentials used by our system.

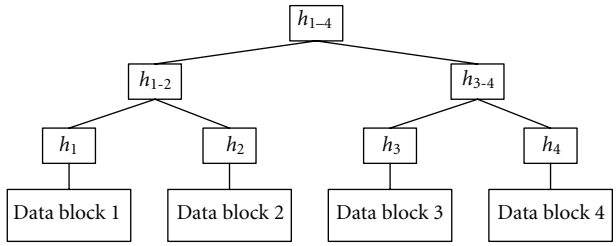


Figure 2: A Merkle hash tree example.

Order-Preserving Symmetric Encryption. The Order-Preserving Symmetric Encryption (OPSE) is a deterministic encryption scheme whose encryption function preserves the numerical order of the plaintexts [24]. More specifically OPSE has the property that given two numbers x and y, $x \leq y$, after encrypted using OPSE function $\mathrm{Enc}(\cdot)$, $\mathrm{Enc}(x) \leq \mathrm{Enc}(y)$. OPSE provides efficient range queries on encrypted data. By "efficient" we mean in time logarithmic in the size of data, as performing linear work on each query is prohibitively slow in practice for a large amount of data. Consider an OPSE function $\mathrm{Enc}(\cdot)$ from plaintext domain $D = 1, \ldots, P$ to ciphertext domain $R = 1, \ldots, Q$, $\mathrm{Enc}(\cdot)$ can be uniquely defined by a combination of P out of Q ordered items. An OPSE algorithm is then said to be secure if and only if an adversary has to perform a brute force search over all possible combinations of P out of Q ordered items to break the encryption scheme. The concept of OPSE was first proposed in database community by Agrawal et al. [25]. In [24] Boldyreva et al. gave the state-of-the-art cryptographic study of OPSE primitive and provided a construction that is provably secure under the security framework of pseudorandom function or pseudorandom permutation. The readers can refer to [24] for more details about OPSE and its security definition.

HMAC. In cryptography, HMAC (Hash-based Message Authentication Code) is a specific construction for calculating a message authentication code (MAC) involving a collision-resistant hash function in combination with a secret key. As with any MAC, it may be used to simultaneously verify both authenticity and integrity of a message. Any cryptographic hash function, such as MD5 or SHA-1, may be used in the calculation of an HMAC. The cryptographic strength of the HMAC depends upon the cryptographic strength of the underlying hash function.

Merkle Hash Tree. Merkle hash tree (MHT) first invented in 1989 by Merkle [26] is a tree of digests in which the leaf is the digest of a data block. Nodes further up in the tree are the digests of their respective children. For example, as shown in Figure 2, $h_{1\text{-}2}$ is the digest of hashing h_1 and h_2. That is, $h_{1\text{-}2} = h(h_0 \| h_1)$ where $\|$ denotes concatenation. Most Merkle hash trees are binary (two child nodes under each node) but they can as well use many more child nodes under each node. The Merkle hash tree is always used to authenticate a set of messages collectively, without authenticating each one individually, for example, [4, 27]. Following the same idea with [4], we propose a novel data structure called AI tree to guarantee the authenticity and integrity of query results from MNs.

Table 1: Primary notations.

Notation	Description
L	Number of SNs in a cell
N	Number of sensed data items in a time slot
n	Number of encrypted data items in a page
m	Number of leaves of upper tree in a page
S	Query result to a query
$h(\mathrm{msg})$	Digest of hashing msg
$\mathrm{Enc}(k, \mathrm{msg})$	OPSE ciphertext of msg
$\mathrm{hmac}(k, \mathrm{msg})$	HMAC of msg
$\lvert P\rvert$, $\lvert d\rvert$	Size of a page and digest
$\lvert e\rvert$	Size of an encrypted data item
$\lvert\mathrm{des}\rvert$, $\lvert\mathrm{lev}\rvert$	Size of a descendant label and level label
height	Height of AI tree
[low, high]	Query range
MHT	Merkle hash tree

Table 1 lists the primary notations used in this paper.

5. The Proposed Scheme

In this section, our scheme, SEF, is described in detail. Without loss of generality, we focus on one cell consisting of L SNs referred to as SN_i, $i = 1, 2, \ldots, L$ and a MN referred to as MN.

5.1. System Initialization. A simple approach for data privacy is to require each SN to encrypt each data item using ordinary symmetric encryption algorithm (such as DES and AES) before sending it to MN. Although this approach leaks the least information to MN and provides strong privacy, it is not suitable for efficient range queries because MN has no knowledge to locate the encrypted data items which exactly match the range. MN has to send the entire encrypted data collected in the specified time slot to the BS. Obviously, this approach is very inefficient and wastes much energy. To address this issue, we employ OPSE in our scheme, which only exposes the order information of the encrypted data to MN. OPSE not only allows efficient range queries, but also maintains a strong security standard.

The BS loads an OPSE function and a hash function to each SN_i ($1 \leq i \leq L$). Additionally the BS loads the hash function to MN. We use the notation $\mathrm{Enc}(k, m)$ for the ciphertext of plaintext m, where the OPSE function is $\mathrm{Enc}(\cdot)$ and the key is k. In order to ensure data privacy against adversaries or compromised MN, the sensed data must be encrypted before sent to MN. So the BS preloads each SN_i with a distinct initial key key_i^0 which is only shared with the BS and SN_i. At the end of time slot t ($t \geq 0$), SN_i computes a new key $\mathrm{key}_i^{t+1} = h(\mathrm{key}_i^t)$, then erases the old key key_i^t. This mechanism has several advantages: (1) compromising SN_i

and MN does not lead to the disclosure of the data values sensed by SN_i before the compromise. (2) As each SN has a distinct key chain, even one SN is compromised, the security of other SNs will not be affected. In addition, the BS chooses two boundary data items, $d_{\min}$ and $d_{\max}$, as the minimum bound and maximum bound, respectively, for all possible sensed data items. For example, the temperature in the range $[-50, 200]$ is considered reasonable in most situations, so the BS can choose $d_{\min} = -100$ and $d_{\max} = 300$. Both $d_{\min}$ and $d_{\max}$ are also loaded into each SN_i $(1 \le i \le L)$.

5.2. Sensor Node Behavior. Sensor node behavior is divided into three phases: sensing phase, AItree-construction phase, and communication phase. In sensing phase, SN_i performs sensing and simple encryption tasks. In AItree-construction phase, on top of all encrypted data items, SN_i builds an AI tree and finally computes the HMAC. In communication phase, SN_i submits all encrypted data items and the HMAC to MN.

In sensing phase, SN_i performs the sensing task first. Assume that SN_i can sense N data items in time slot t. For each sensed data item d_j, $1 \le j \le N$, d_j is encrypted into $e_j = \text{Enc}(\text{key}_i^t, d_j)$. In addition, SN_i computes $e_0 = \text{Enc}(\text{key}_i^t, d_{\min})$ and $e_{N+1} = \text{Enc}(\text{key}_i^t, d_{\max})$ as the minimum bound and maximum bound for all e_j, $1 \le j \le N$.

Since the I/O interface of NAND flash does not provide a random-access external address bus, data must be read on a pagewise basis, our scheme makes full use of this characteristic to achieve high storage utilization and query processing efficiency. In AItree-construction phase, on top of e_j, $0 \le j \le N+1$, SN_i constructs an AI tree which consists of an upper tree and some lower trees as illustrated in Figure 3. First of all, SN_i sorts e_j, $0 \le j \le N+1$. Then SN_i partitions the $(N+2)$ encrypted data items into pages, each page contains n encrypted data items except the last page which contains $(N+2)\%n$ encrypted data items. Each page should satisfy the condition:

$$|e| \times n \le |P|, \tag{2}$$

where $|e|$, $|P|$ are the size of an encrypted data item and a page of NAND flash, respectively. In order to achieve the highest storage utilization, n is set to the largest value that satisfies (2). Based on each page, a lower tree is constructed. Afterward an upper tree is constructed based on the roots of the lower trees. Finally SN_i computes the HAMC using the root of the AI tree.

In order to record the structure of the AI tree, two labels are attached to each digest, called descendant label and level label. We use the notation $h_{\text{level}}^{\text{des}}$ to represent the digest and these two labels. When des is set to "left", it indicates the digest is the left child of its parent; while des is set to "right", the digest is the right child of its parent. *level* is an integer indicating which level the digest is in the AI tree. If the height of the AI tree is *height*, height ≥ 0, the *level* of the root is *height*. And the *level* of a child node is one less than that of its parent. Since the root of the AI tree has no sibling, des of the root is set to "left" or "right" is inessential.

For visualization, here we use an example to describe the process. Figure 3 illustrates an AI tree in a scenario where there are $N = 9$ data items (except e_0 and e_{10}) sensed by SN_i in time slot t. The 11 data items are sorted in ascending order. Each page can contain 4 data items. The leaf of the AI tree is the encrypted data item; the internal nodes and the root are the digests of their respective child nodes. Here $\text{HMAC} = \text{hmac}(\text{key}_i^t, h_{0-10})$.

In order to minimize the energy consumption, SN_i does not send the entire AI tree to MN. Our scheme is based on the observation that the energy consumption of communication is significantly more than that of hash operation. In order to verify this observation, we make a experiment on TelosB motes [28]. Figure 4 presents the energy consumption of communication and hash operation. Note that the communication involves two parties: the sender and the receiver; the energy consumption of communication is the sum of the sender's and receiver's energy consumption. It can be concluded that the consumption of communication is much more than that of hash operation. So in communication phase, SN_i only sends the HMAC and encrypted data items (e_0 to e_{10}) to MN. MN rebuilds the AI tree using the encrypted data items received. Though rebuilding the AI tree consumes a little energy, this approach reduces a lot of communication cost, so reduces the total energy consumption. More importantly, as SN is energy limited but MN has more energy supply, by transferring the communication task between SN and MN to the hash task at MN, this approach significantly prolongs the network's life cycle.

5.3. Master Node Behavior. After receiving all encrypted data items and HMAC, MN rebuilds the AI tree, then stores all encrypted data items, HMAC and partial digests of the AI tree in NAND flash. Considering the characteristic of NAND flash, that is, NAND flash is divided into pages, in Section 5.5, we will discuss which digests of the AI tree are stored, and how to store these digests.

If the BS wants to query MN for the data items in range [low, high] sensed by SN_i in time slot t, it launches a query request specified by $[\text{Enc}(\text{key}_i^t, \text{low}), \text{Enc}(\text{key}_i^t, \text{high})]$. Upon receiving the range query request, MN returns the query result S to the BS. S consists of all encrypted data items in $[\text{Enc}(\text{key}_i^t, \text{low}), \text{Enc}(\text{key}_i^t, \text{high})]$ and two boundary data items, $B_{\min}$ and $B_{\max}$, which fall immediately to the minimum and to the maximum of the query range. The inclusion of $B_{\min}$ and $B_{\max}$ is to confirm that all qualified data items are contained by S. Additionally MN returns to the BS a verification object referred to as VO, which is used to verify the authenticity and integrity of S by the BS. VO consists of three components: (1) the HMAC, (2) all left sibling digests to the path of $h(B_{\min})$, and (3) all right sibling digests to the path of $h(B_{\max})$. After receiving S and VO from MN, the BS makes use of S and component (2) and (3) of VO to compute HMAC. If the computed HAMC is equal to component (1) of VO, S is definitely both authentic and integral. Otherwise S is incorrect.

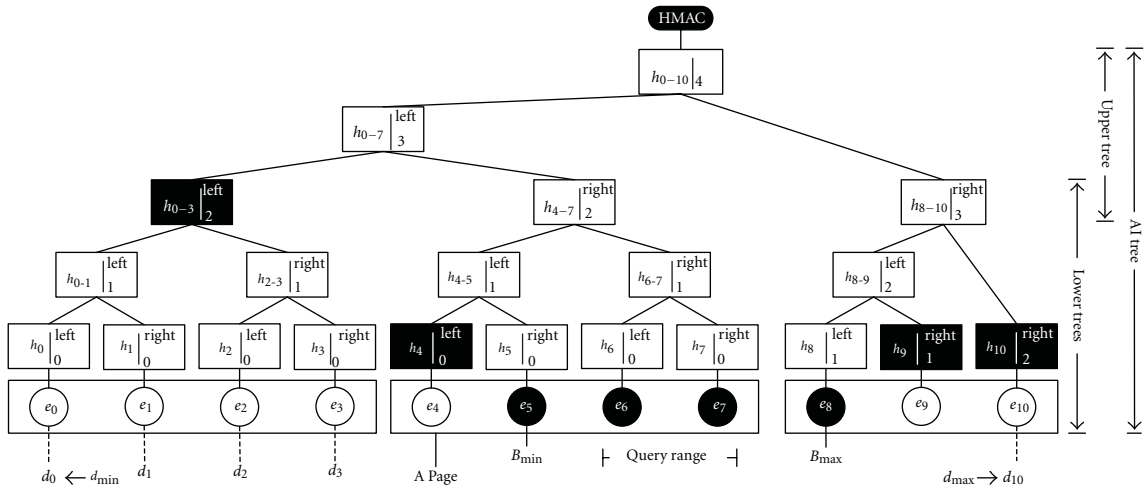


Figure 3: An AI tree example.

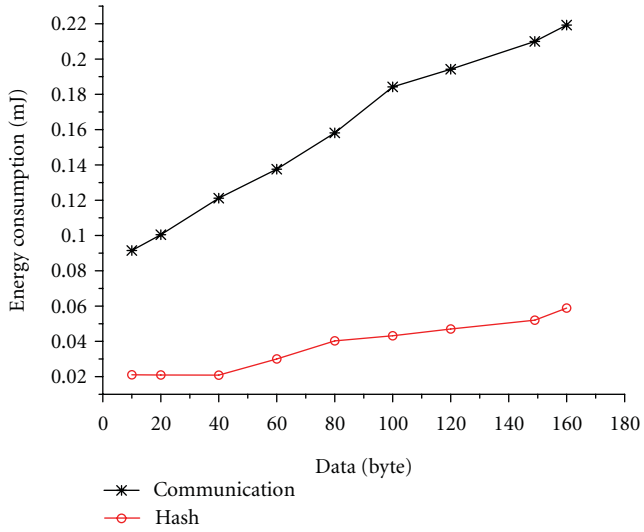


Figure 4: Average energy consumption of communication and hash.

For example, suppose the BS wants to query the data sensed by SN_i in time slot t, and the query range is $[e_6, e_7]$. As shown in Figure 3, since e_5 and e_8 fall immediately to the query range, the boundary data items are $B_{\min} = e_5$ and $B_{\max} = e_8$. The paths leading to $h(B_{\min})$ and $h(B_{\max})$ are shown bold. $S = \{e_5, e_6, e_7, e_8\}$. Component (1) of VO is the HMAC = hmac$(\text{key}_i^t, h_{0-10})$. Component (2) of VO consists of all left sibling digests to the path of h_5, that is, h_{0-3} and h_4. Component (3) of VO consists of all right sibling digests to the path of h_8, that is, h_{10} and h_9. Here we omit the descendant labels and level labels attached to digests. Note that the AI tree in Figure 3 is not a full binary tree, h_{10} is actually the right sibling of $h_{8\text{-}9}$.

In a real-world scenario not all applications require the authenticity and integrity guarantee. For example, some resource-critical applications may favor an energy-saving response over a verified one. In these situations, MN only returns the query result S, achieving the performance of a nonauthenticated range query scheme. But in some cases the authenticity and integrity guarantee is required, MN returns VO in addition to S. By keeping encrypted data separated from verification information, our scheme provides much flexibility and optimizes the performance of applications. Though the MHT needs not be binary, that is, it can be a multiway tree, the AI tree must be binary to minimize the number of digests in VO. Suppose the AI tree is an x-way tree (i.e., each parent node has x child nodes), the maximal number of digests in VO is $f(x) = (x-1)\log_x(N+2)$. Clearly $f(x)$ is a strictly monotonically increasing function. So only $x = 2$, the size of VO is minimal.

5.4. Network Owner Behavior. After receiving VO and S, the BS reconstructs the AI tree and computes HMAC, then verifies that the computed HMAC is equal to component (1) of VO. The details of the process is described as follows.

The BS uses a digest array to store the temporal digests computed on the fly. For simplicity of expression, we call the digest inserted into the digest array first the most left digest denoted by h^l, and that inserted into the digest array last is the most right digest denoted by h^r. First of all, the BS hashes all encrypted data items in S, and stores these digest in the digest array. On the level η, (η from 0 to *height*), according to the number of digests inserted into VO, there are three cases.

(i) There are two digests to which the level label η is attached, denoted by h_η^{des} (des is "left" and "right"). The BS computes $h(h^l \| h_\eta^{\text{left}})$ and $h(h_\eta^{\text{right}} \| h^r)$ to derive the parent digests.

(ii) There is only one digest to which the level label η is attached, denoted by h_η^{des} (des is "left" or "right"). If des is "left", the BS computes $h(h^l \| h_\eta^{\text{left}})$ to derive the parent digest. Otherwise the BS computes $h(h_\eta^{\text{right}} \| h^r)$ to derive the parent digest.

(iii) There is no digest to which the level label η is attached, the BS takes no action.

Then, the BS processes the unprocessed digests (including h^l and h^r in the (iii) case). These unprocessed digests are concatenated pairwise from left to right to derive the parent digests. As the AI tree may not be a full binary tree, the last digest in the digest array may have no sibling. In this case, the last digest is not processed temporarily. This process is carried out until there is only one digest left in the digest array, which is exactly the root of the AI tree.

Recalling the example in Figure 3, the BS wants to query the data sensed by SN_i in time slot $t, S = \{e_5, e_6, e_7, e_8\}$ and VO $= \{h_{0-3}, h_4, h_9, h_{10}\}$. During verification, the BS concatenates h_4 and h_5 to calculate $h_{4\text{-}5}$, h_6 and h_7 to calculate $h_{6\text{-}7}$. Then the BS appends $h_{6\text{-}7}$ to $h_{4\text{-}5}$ to derive h_{4-7}. After that it appends h_{4-7} to h_{0-3} to derive h_{0-7}. Similarly, it calculates h_{8-10}. Finally, BS gets $h_{0-10} = h(h_{0-7} \parallel h_{8-10})$, and verifies whether $\text{hmac}(\text{key}_i^t, h_{0-10})$ is equal to component (1) of VO.

5.5. AI Tree Compression. To reduce the storage overhead and the cost of reading data from NAND flash, MN does not materialize the entire AI tree. That is some internal nodes in AI tree will not be stored in MN, but recomputed on the fly when necessary. Since the I/O interface of NAND flash does not provide a random-access external address bus, data must be read on a pagewise basis, our scheme makes full use of this characteristic to achieve the optimal storage utilization and query processing efficiency.

Taking into account the characteristic of NAND flash, with respect to the lower trees, MN only stores the leaves of the lower trees (i.e., the encrypted data items). For example, in Figure 3, when the BS wants to query e_6 and e_7, MN has to load the entire page, so e_4 and e_5 are also loaded, then MN computes h_4, h_5, h_6, and h_7 and finally gets h_{4-7}. All these digests are computed on the fly.

With regard to the upper tree, a basic approach is that only the leaves of the upper tree are stored. The leaves of the upper tree are read from NAND flash immediately, and the internal nodes of the upper tree can be computed on the fly using the leaves of the upper tree. For instance, in Figure 3, assume that a page of NAND flash can accommodate more than 3 digests, we use a single page to store h_{0-3}, h_{4-7} and $h_{9\text{-}10}$. If h_{0-7} is a part of VO, MN only needs to load this page and computes $h_{0-7} = h(h_{0-3} \parallel h_{4-7})$ on the fly. Though this approach introduces little storage overhead, it is not efficient if there is a lot of data, because the cost of reading leaves of the upper tree from NAND flash may be expensive. For instance, if there are p pages storing the leaves of the upper tree, but the BS only wants to query a small fraction of the sensed data, the roots of which are placed in p^* pages, MN has to read at least $p - p^*$ pages to get the other leaves of the upper tree.

To address the above issue, we apply the optimized approach of [4]. Motivated by the characteristic of NAND flash, that is, reading data from NAND flash is on a pagewise basis, MN stores all digests of the upper tree required to construct VO in one page. Following the same idea with lower trees, MN also partitions the upper tree leaves (i.e., the lower tree root) into pages, suppose each page can accommodate m upper tree leaves. For each page, MN builds a binary compressed MHT. Afterward, a cusp tree is constructed on top of all MHTs' roots. For each MHT, only the leaves and sibling digests to the path of the MHT's root are materialized; other nodes are not materialized, they are computed on the fly when required. As shown in Figure 5, the black rectangles represent the materialized nodes. Those not materialized are represented by the striped rectangles. Here we omit the descendant and level labels attached to each digest and the unrelated nodes. Since data must be read on a pagewise basis, in order to reduce the cost of loading the upper tree, we place the m leaves and the sibling digests to the path of the MHT's root (h_1 and h_2) in one page of NAND flash. The number of sibling digests to the path of the MHT's root is equal to the height of the cusp tree, which is $\lceil \log(\lceil (\lceil N/n \rceil)/m \rceil) \rceil$ ($\lceil\ \rceil$ means round up). Each page should satisfy the condition:

$$\left(m + \left\lceil \log\left(\left\lceil \frac{(\lceil N/n \rceil)}{m} \right\rceil\right) \right\rceil\right) \times (|d| + |\text{des}| + |\text{lev}|) \leq |P|, \tag{3}$$

where $|d|$, $|\text{des}|$, $|\text{lev}|$, $|P|$ are the sizes of a digest, a descendant label, a level label, and a page of NAND flash, respectively. In order to minimize the storage overhead, we set m to the largest value that satisfies (3).

As reading data from NAND flash is in pages, when one page is loaded, all digests of the upper tree required to construct VO are loaded. If $B_{\min}$ and $B_{\max}$ are covered by two different MHTs, two page accesses are required; otherwise, only one page access is required.

But there exists an extreme case in which N is so huge that m is equal to 0. We define the page utilization ratio as

$$\delta = \frac{m}{\lceil \log(\lceil (\lceil N/n \rceil)/m \rceil) \rceil}. \tag{4}$$

Given δ, we use λ pages to store these digests. That is,

$$m + \left\lceil \log\left(\left\lceil \frac{(\lceil N/n \rceil)}{m} \right\rceil\right) \right\rceil \times (|d| + |\text{des}| + |\text{lev}|) \leq \lambda |P|. \tag{5}$$

m is set to the maximum that satisfies (4) and (5). λ is set to the minimum that satisfies (5). In this case, if $B_{\min}$ and $B_{\max}$ are covered by two different MHTs, 2λ page accesses are required; otherwise, only λ page accesses is required.

6. Security Analysis

6.1. Privacy Analysis. In a two-tiered WSN protected by our scheme, even if an arbitrary number of MNs is compromised, the adversary cannot obtain the actual values of data items collected by SNs and the actual queries issued by the BS. The reasons are given as follows. Both in the submission and in the query, the data items and the query range sent to the MNs are all encrypted using order-preserving encryption algorithm. Without knowing the keys used in the encryption, it is computationally infeasible for the adversary to get the actual values of data items and query range.

If a SN is compromised, the collected data and key are exposed to adversaries, and the compromised SN may send

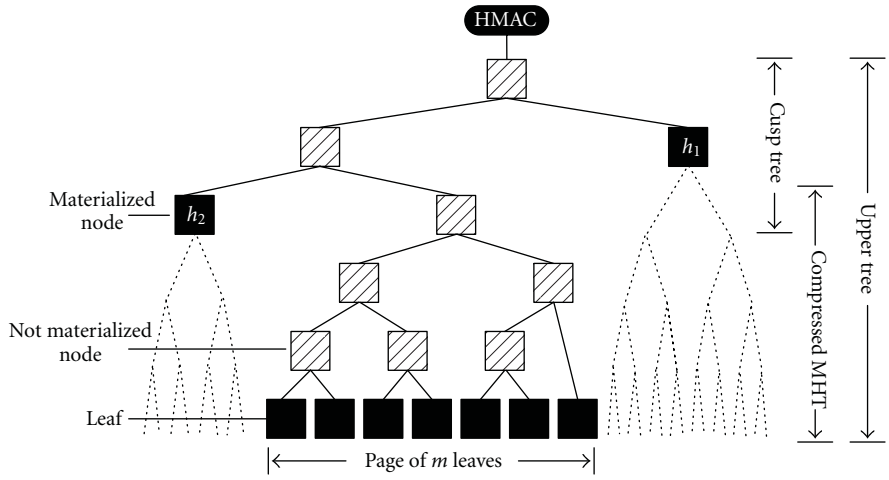


Figure 5: An upper tree compression example.

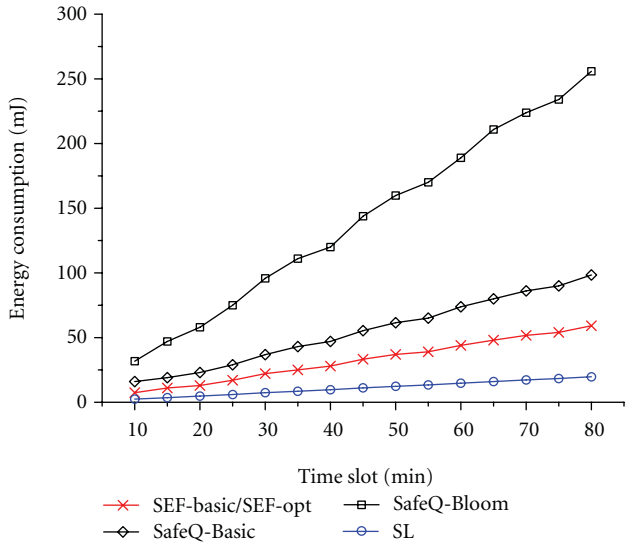


Figure 6: Average energy consumption of processing data for a SN.

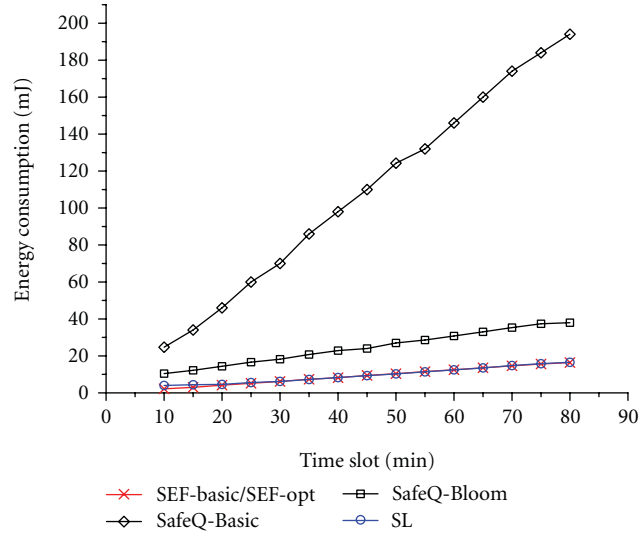


Figure 7: Average energy consumption of submission.

forged data to its MN; it is hard to prevent such an attack without the use of tamper proof hardware. In our scheme, the key is updated every time slot and the old keys are erased, even a SN and its MN are compromised, adversaries cannot obtain the data values before the compromise. Additionally, as each SN shares a distinct key with the BS, even an SN is compromised, other SNs will not be affected.

6.2. Integrity Analysis. Let $[e_{\text{low}}, e_{\text{high}}]$ be the query range, $S^* = \{e_{n_1^*}, \ldots, e_{i^*}, \ldots, e_{n_2^*}\}$ be the correct query result and $S = \{e_{n_1}, \ldots, e_i, \ldots, e_{n_2}\}$ be the query result from MN. Here $e_{n_1^*} = B_{\min}$ and $e_{n_2^*} = B_{\max}$. The BS can verify that MN has not inserted any forged data item or excluded, tampered any qualified data item.

(1) If there exists $n_1^* \le i^* \le n_2^*$ such that $e_i^* \notin S$ (that is a qualified data item is deleted) or e_i^* is tampered, or if there exists $n_1 \le i \le n_2$ such that $e_i \notin S^*$ (that is a forged data item is inserted), the BS can detect this error. Because the structure of AI tree constructed by the BS using S and VO are different from the structure of the AI tree constructed by SN_i. The collision resistance of the hash function guarantees that adversaries cannot insert, delete or tamper any data item in a way leads to an identical root of the AI tree.

(2) Since $B_{\min}$ and $B_{\max}$ fall immediately to the query range, if $n_1 > n_1^*$, the BS can detect this error because there is no item in S less than e_{low}. If $n_1 < n_1^*$, the BS can also detect this error, because there are more than one item in S less than e_{low}. The same as n_1, if $n_2 > n_2^*$, there is more than one item in S larger than e_{high}; and if $n_2 < n_2^*$, there is no item in S larger than e_{high}. The use of $B_{\min}$ and $B_{\max}$ ensures that all qualified encrypted data items are included in S.

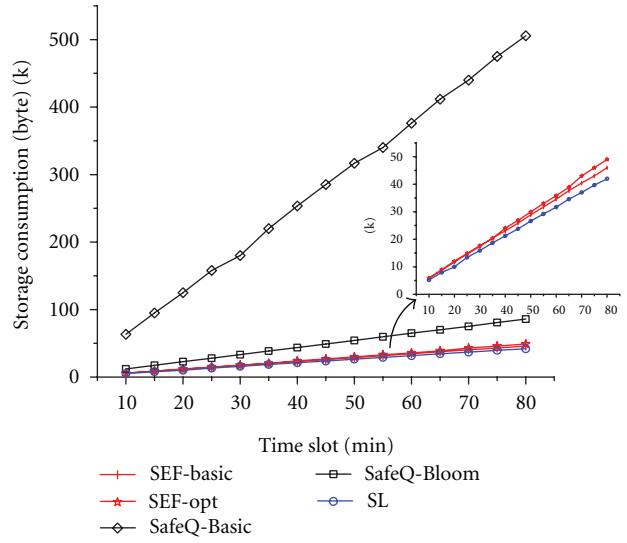


Figure 8: Average storage consumption for MN.

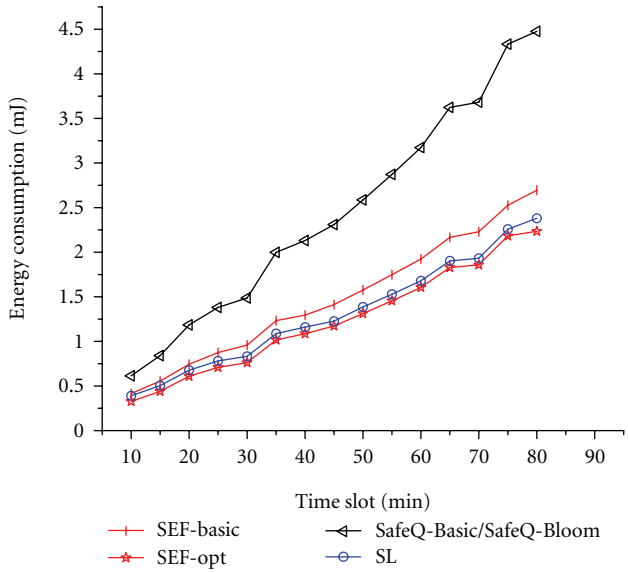


Figure 9: Average energy consumption of query processing for MN.

(3) If e_{low} (e_{high}) is the minimum (maximum) for all possible data items, the lower bound (upper bound) for all possible data items, d_{min} (d_{max}), takes the role of the boundary value, that is, $B_{min} = E(d_{min})$ ($B_{max} = E(d_{max})$). These two boundary values, d_{min} and d_{max}, guarantee that any query range can be processed correctly.

7. Implementation and Performance Evaluation

In this section, we evaluate the performance of our scheme and perform side-by-side comparison with SL scheme SafeQ-Basic, and SafeQ-Bloom in a test-bed. We analyze the performance comparison in five aspects: the energy consumption of processing sensed data for an SN; the energy consumption of submission from an SN to MN; the storage

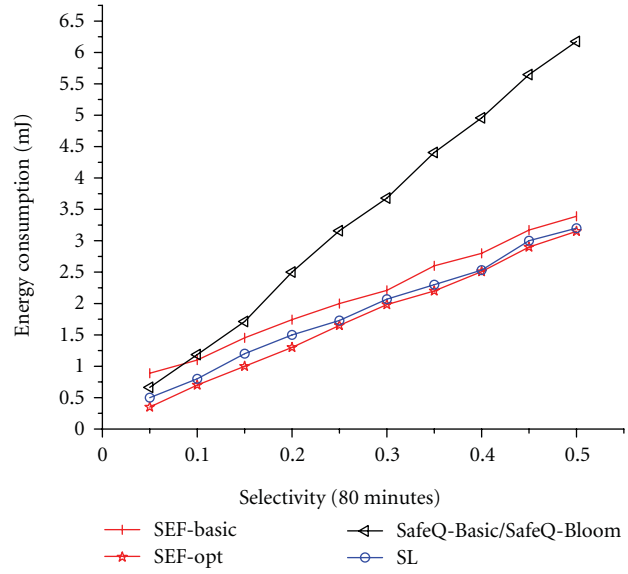


Figure 10: Average energy consumption of query processing for MN.

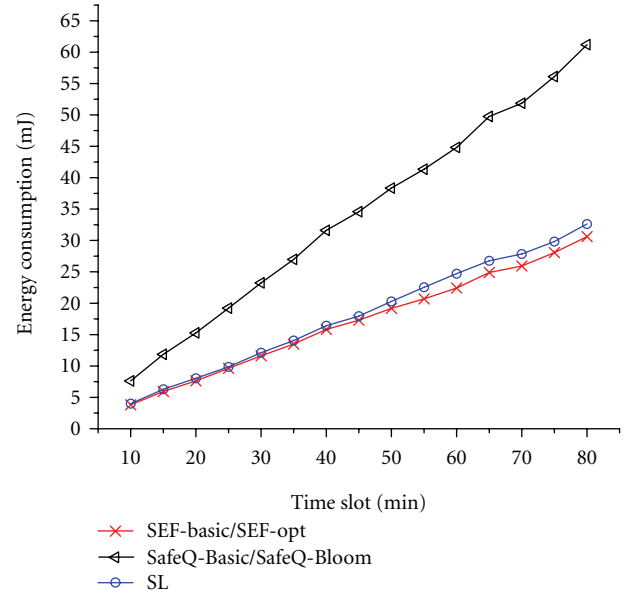


Figure 11: Average energy consumption of sending response for MN.

consumption for MN; the energy consumption of query processing for MN; the energy consumption of sending response to a query to the BS for MN. As the BS is a powerful device, the cost for the BS is neglectable. In our experiments, we do not choose the schemes of [9, 10] for side-by-side comparison for two reasons. First the techniques used in [9, 10] are similar to [8] except the optimization for integrity verification. Second the optimization incurs serious security problems, a compromised SN can easily breach the integrity verification of the network by sending false bit maps.

7.1. Experimental Test Bed and Setup. The test-bed consists of eleven TelosB motes as SNs and one TelosB mote as

MN. Note that the number of SNs in a cell is application dependent. If the number of SNs is large, the storage or the energy of MN is exhausted quickly and reduces the life cycle of the WSN. While, if the number of SNs is small, the efficiency of the WSN is very low. SNs should be carefully deployed to achieve the best equilibrium. The TelosB mote is equipped with an 8 MHz CPU, 10 KB RAM, 48 KB ROM, 1024 KB flash, and an 802.15.4/ZigBee radio [28]. The flash of the TelosB mote is divided into pages of 256 bytes. These TelsoB motes run on the operating system Tinyos-2.1.0 [29].

We use SHA-1 [30] as the hash function. Note that it has been pointed out that there are collision attacks on SHA-1 [31]. However, since it requires about 2^{69} hash operations, it is difficult to launch the attacks in practice [32]. The HMAC operation also uses SHA-1 as the underlying hash function, both the size of a digest and an HMAC is 20 bytes. The size of an encrypted data item is set to 4 bytes. An 8-bits array is used to represent the descendant label and level label, which can record as much as 2^{128} encrypted data items. We use the DES encryption algorithm in implementing SL scheme, SafeQ-Basic and SafeQ-Bloom as in [8, 11, 12]. We employ the OPSE function proposed in [24] as the underlying OPSE function.

To present which portion of the data are queried, we introduce the query *selectivity*. Assume that the number of all possible sensed data items is Z, the accuracy of sensed data is a. If the BS wants to query the data items in range [low, high], selectivity $= ((\text{high} - \text{low})/a + 1)/Z$. For example, here we suppose the values of sensed data are integers, the number of all possible sensed data items $Z = 100$, and the query range [low, high] $= [25, 50]$. In this case, the selectivity should equal $(50 - 25 + 1)/100$.

As in [8], we use their optimal bucket partition algorithm for computing the optimal bucket partition in implementing SL scheme. Note that in some situations, it is hard to get the accurate distribution of sensed data in advance, the BS is unable to obtain the optimal bucket partition. This factor will greatly influence the performance of SL scheme. We use SEF-basic and SEF-opt to denote our schemes with basic and optimized AI tree compression, respectively.

7.2. Result Summary

(1) Figure 6 depicts the energy consumption of processing sensed data for an SN. Our experiments show that, compared with SL scheme, SEF-basic/SEF-opt consume 3 times more energy since it has to perform some operations (such as hash operation) to construct the AI tree. But they have great advantage compared with SafeQ-Basic and SafeQ-Bloom. Compared with SEF-basic/SEF-opt, SafeQ-Basic consumes 1.67 times more energy and SafeQ-Bloom consumes 4.3 times more energy, because SafeQ-Basic and SafeQ-Bloom adopt the prefix membership verification scheme which requires a large number of hash operations. As SafeQ-Bloom has to perform a lot of additional hash operations to obtain the Bloom filter, it consumes the most energy.

(2) Figure 7 shows the energy consumption of submitting data from an SN to MN. Note that the submission involves two parties: the SN and MN, so the energy consumption is the total energy consumption of the SN and MN. Our experiments show that SL scheme performs almost as well as SEF-basic/SEF-opt. In comparison with SEF-basic/SEF-opt, SafeQ-Basic consumes **11.8** times more energy and SafeQ-Bloom consumes **2.5** times more energy. Due to a mass of HMACs produced by prefix membership verification scheme, SafeQ-Basic pays a lot of energy on submission.

(3) Figure 8 illustrates the storage consumption for MN. As we can see that the storage consumption of SEF-basic and SEF-opt is very close to that of SL scheme. Compared with SL scheme, SEF-opt incurs only 11% storage overhead on average. And compared with SEF-opt, SafeQ-Basic requires up to **10.6** times more storage. Though employing the Bloom filter technique, SafeQ-Bloom still consumes 1.8 times more storage compared with SEF-opt.

(4) We carry out two kinds of experiments to measure the energy consumption of query processing for MN. One is to vary the time slot and generate range queries randomly. The other is to fix the time slot (we set the time slot to 80 minutes) and vary *selectivity*. Given a *selectivity*, *low* is generated randomly.

In the first case, as shown in Figure 9, SEF-opt consumes the least energy, it saves 21% energy compared with SEF-basic and 6.7% energy compared with SL scheme. SafeQ-Basic/SafeQ-Bloom is always consuming the most energy. Compared with SEF-opt, they consume as much as 2 times more energy. When time slot is 80 minutes, SL scheme consumes 2.37 mJ energy, SafeQ-Basic/SafeQ-Bloom consumes 4.47 mJ energy, SEF-basic consumes 2.69 mJ energy, while SEF-opt consumes only 2.23 mJ energy.

In the second case, as shown in Figure 10, SEF-opt is overall the most efficient scheme for any ratio. With the increase of selectivity, the energy consumption of SEF-basic is close to that of SEF-opt more and more. When the *selectivity* is 0.1, SEF-opt consumes the least energy, 0.7 mJ, SEF-basic consumes 1.1 mJ energy, SafeQ-Basic/SafeQ-Bloom consumes 1.18 mJ energy, SL scheme consumes 0.8 mJ energy. It is apparent that SafeQ-Basic/SafeQ-Bloom consumes the most energy except *selectivity* is less than 0.07. Compared with SEF-opt, SafeQ-Basic/SafeQ-Bloom consumes nearly 2 times more energy.

(5) Figure 11 presents the energy consumption of sending response to a query to the BS for MN. The response includes the query result and verification information. Since the false positive problem incurred by bucketing technique, the performance of SL scheme is a little worse than that of SEF-basic/SEF-opt. Both our schemes consume about 2 times less energy than SafeQ-Basic/SafeQ-Bloom.

8. Conclusions

In this paper, we have proposed a secure, efficient, and flexible scheme, called SEF, for range queries in two-tiered WSNs. Our scheme employs the order-preserving symmetric encryption algorithm to preserve data privacy. We have proposed a novel data structure called AI tree to guarantee the authenticity and integrity of query results. We first

consider the storage media issue in the design of range query schemes and present how to use the characteristic of NAND flash to achieve high storage utilization and query processing efficiency. The experiments have shown that SEF significantly achieves efficiency.

Acknowledgments

This work is supported by the National Science Foundation of China under Grants no. 61070155 and no. 60903153, Program for New Century Excellent Talents in University (NCET-09-0685), the Fundamental Research Funds for the Central Universities (DUT10ZD110), and the SRF for ROCS, SEM.

References

[1] S. Shenker, S. Ratnasamy, B. Karp, R. Govindan, and D. Estrin, "Datacentric storage in sensornets," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 1, pp. 137–142, 2003.

[2] P. Desnoyers, D. Ganesan, H. Li, and P. Shenoy, "Presto: a predictive storage architecture for sensor networks," in *Proceedings of the 10th Workshop on Hot Topics in Operating Systems (HotOS X)*, June 2005.

[3] F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin, "Dynamic authenticated index structures for outsourced databases," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pp. 121–132, June 2006.

[4] K. Mouratidis, D. Sacharidis, and H. Pang, "Partially materialized digest scheme: an efficient verification method for outsourced databases," *The Very Large Data Bases Journal*, vol. 18, no. 1, pp. 363–381, 2009.

[5] H. Pang, J. Zhang, and K. Mouratidis, "Scalable verification for outsourced dynamic databases," in *Proceedings of the Very Large Data Bases Conference (VLDB)*, pp. 802–813, 2009.

[6] D. He, Y. Gao, S. Chan, C. Chen, and J. Bu, "An enhanced two-factor user authentication scheme in wireless sensor networks," *Ad Hoc & Sensor Wireless Networks*, vol. 10, no. 4, pp. 361–371, 2010.

[7] D. He, J. Bu, S. Zhu et al., "Distributed privacy-preserving access control in a single-owner multi-user sensor network," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM) mini-conference*, 2011.

[8] B. Sheng and Q. Li, "Verifiable privacy-preserving range query in two-tiered sensor networks," in *Proceedings of the 27th IEEE Communications Society Conference on Computer Communications, IEEE INFOCOM 2008*, pp. 457–465, April 2008.

[9] J. Shi, R. Zhang, and Y. Zhang, "Secure range queries in tiered sensor networks," in *Proceedings of the 28th IEEE Conference on Computer Communications, IEEE INFOCOM 2009*, pp. 945–953, Rio de Janeiro, Brazil, April 2009.

[10] R. Zhang, J. Shi, and Y. Zhang, "Secure multidimensional range queries in sensor networks," in *Proceedings of the 10th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc*, pp. 197–206, May 2009.

[11] F. Chen and A. X. Liu, "SafeQ: secure and efficient query processing in sensor networks," in *Proceedings of the 29th IEEE Conference on Computer Communications, IEEE INFOCOM 2010*, pp. 1–9, East Lansing, Mich, USA, March 2010.

[12] F. Chen and A. X. Liu, "Privacy and integrity preserving range queries in sensor networks," Tech. Rep. MSU-CSE-09-26, Michigan State University, Ann Arbor, Mich, USA, 2010.

[13] B. Hore, S. Mehrotra, and G. Tsudik, "A privacypreserving index for range queries," in *Proceedings of the Very Large Data Base Conference*, pp. 720–731, Toronto, Canada, 2004.

[14] D. Comer, "Ubiquitous b-tree," *ACM Computing Surveys*, vol. 11, no. 2, pp. 121–137, 1979.

[15] J. Cheng, H. Yang, S. H. Y. Wong, P. Zerfos, and S. Lu, "Design and implementation of cross-domain cooperative firewall," in *Proceedings of the 15th IEEE International Conference on Network Protocols, ICNP 2007*, pp. 284–293, Beijing, China, October 2007.

[16] S. Nath and A. Kansal, "FlashDB: dynamic self-tuning database for NAND flash," in *Proceedings of the IPSN 2007: 6th International Symposium on Information Processing in Sensor Networks*, pp. 410–419, New York, NY, USA, April 2007.

[17] G. Mathur, P. Desnoyers, P. Chukiu, D. Ganesan, and P. Shenoy, "Ultra-low power data storage for sensor networks," *ACM Transactions on Sensor Networks*, vol. 5, no. 4, 2009.

[18] K. Ren, W. Lou, and Y. Zhang, "LEDS: providing location-aware end-to-end data security in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 5, Article ID 4358997, pp. 585–598, 2008.

[19] Y. Zhou and Y. Fang, "A two-layer key establishment scheme for wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 9, pp. 1009–1020, 2007.

[20] Q. Wang, K. Ren, W. Lou, and Y. Zhang, "Dependable and secure sensor data storage with dynamic integrity assurance," in *Proceedings of the 28th Conference on Computer Communications, IEEE INFOCOM 2009*, pp. 954–962, Rio de Janeiro, Brazil, April 2009.

[21] Y. Zhou, Y. Zhang, and Y. Fang, "Access control in wireless sensor networks," *Ad Hoc Networks*, vol. 5, no. 1, pp. 3–13, 2007.

[22] W. Du, Y. S. Han, J. Deng, and P. K. Varshney, "A pairwise key pre-distribution scheme for wireless sensor networks," in *Proceedings of the 10th ACM Conference on Computer and Communications Security, CCS 2003*, pp. 42–51, October 2003.

[23] D. He, L. Cui, H. Huang, and M. Ma, "Design and verification of enhanced secure localization scheme in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 7, pp. 1050–1058, 2009.

[24] A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, "Order-preserving symmetric encryption," *Advances in Cryptology—EUROCRYPT 2009*, vol. 5479, pp. 224–241, 2009.

[25] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2004*, pp. 563–574, June 2004.

[26] R. C. Merkle, "A certified digital signature," in *Advances in Cryptology (CRYPTO'89)*, Lecture Notes in Computer Science, pp. 218–238, Springer, New York, NY, USA, 1989.

[27] H. Pang and K. L. Tan, "Authenticating query results in edge computing," in *Proceedings of the 20th International Conference on Data Engineering—ICDE 2004*, pp. 560–571, April 2004.

[28] http://www.willow.co.uk/TelosB_Datasheet.pdf.

[29] http://docs.tinyos.net.

[30] National institute of standards and technology, U.S. department of commerce, secure hash standard, U.S. federal information processing standard publication, 2002.

[31] X. Wang, Y. L. Yin, and H. Yu, "Finding collisions in the full SHA-1," in *Advances in Cryptology (CRYPTO'05)*, V. Shoup, Ed., vol. 3621 of *Lecture Notes in Computer Science*, pp. 17–36, Springer, New York, NY, USA, 2005.

[32] "IAIK Krypto group-description of SHA-1 collision search project," http://www.iaik.tugraz.at/content/research/krypto/sha1/SHA1CollisionDescription.php.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 282986, 10 pages
doi:10.1155/2011/282986

Research Article

A Biometric Key Establishment Protocol for Body Area Networks

Lin Yao,[1] Bing Liu,[1] Guowei Wu,[1] Kai Yao,[2] and Jia Wang[1]

[1] *School of Software, Dalian University of Technology, Dalian 116023, China*
[2] *Library Information Center, Shenyang University of Technology, Shenyang 110178, China*

Correspondence should be addressed to Guowei Wu, wgwdut@dlut.edu.cn

Received 14 March 2011; Accepted 19 May 2011

Current advances in semiconductor technology have made it possible to implant a network of biosensors inside the human body for health monitoring. In the context of a body area network (BAN), the confidentiality and integrity of the sensitive health information is particularly important. In this paper, we present an ECG (electrocardiogram)-signal-based key establishment protocol to secure the communication between every sensor and the control unit before the physiological data are transferred to external networks for remote analysis or diagnosis. The uniqueness of ECG signal guarantees that our protocol can provide long, random, distinctive and temporal variant keys. Biometric Encryption technique is applied to achieve the mutual authentication and derive a non-linkable session key between every sensor and the control unit. The correctness of the proposed key establishment protocol is formally verified based on SVO logic. Security analysis shows that our protocol can guarantee data confidentiality, authenticity and integrity. Performance analysis shows that it is a lightweight protocol.

1. Introduction

Current advances in semiconductor technology have made it possible to implant some sensor nodes inside the human body for health monitoring. In such a deployment scenario of sensors, some wearable sensors called biometric sensors or biosensors will be usually located on the body surface to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion, or pollutants. These biosensors form a wireless network between themselves, which is called the body area network (BAN) [1–3]. Before the physiological data detected are passed on to external networks for remote analysis, diagnosis, or treatment, these data must be transferred to a single body control unit (CU) such as the sink node worn on the human body. Thus, data transmission can be classified into three levels [4]: (1) between the CU and every biosensor in the BAN; (2) between the CU and a remote server; (3) between the remote server and the physicians.

While the communication rate specifications in a BAN are typically low, the security requirements are stringent, especially when sensitive medical data are exchanged. It should be impossible for an adversary to eavesdrop, inject, and modificate these sensitive data. Privacy laws and regulations also mandate that security and privacy must be guaranteed when the patient-related data are created, transferred, stored, and processed in the BAN [5]. On one hand, authentication between a biosensor and CU must be performed over the network. Without the successful authentication, an attacker may pretend to be a user and transmit false data to the CU, which may lead a wrong diagnosis. On the other hand, data encryption is also important to prevent an attacker from sniffering or modificating these sensitive data. Therefore, a common session key is needed to secure the communication.

There are two classical personal authentication approaches in traditional cryptosystems: (1) knowledge-based approach; (2) token-based approach. Token or knowledge is prone to be forgotten, lost, stolen, or duplicated. Either approach cannot represent the unique user. Compared with knowledge-based approach and token based approach, biometrics can represent the uniqueness of one user. However, the conventional biometric characteristics such as iris, fingerprint, or face are static biometrics. A novel kind of biometric traits, such as heart rate variability, interpulse interval, and other features of electrocardiogram and photoplethysmogram, has been recently studied as a new identification approach [4]. In our protocol, ECG as the dynamic biometrics is utilized to authenticate between a biosensor and CU. The idea of using ECG comes from the observation that the human body is dynamic and complex and the physiological state of a subject is quite randomness and time

variance [6]. ECG is typically collected and utilized in many recognition applications, which is in accordance with the data minimization principle [7].

In order to ensure data confidentiality and privacy, cryptographic algorithms are classified into symmetric and asymmetric algorithms. Asymmetric algorithms require more computation resources and storage resources compared with symmetric algorithms. Asymmetric key or public key cryptography is unsuitable in BAN, because sensor nodes are limited in power, computational capacities, and memory. In our protocol, we prefer the symmetric cryptographic algorithm. Based on the Biometric Encryption [8, 9], a symmetric cryptographic key is established between every biosensor and CU.

In our previous work, we have proposed a key establishment protocol, named by ESKE, based on ECG signals combining with the fuzzy vault scheme [10]. When a CU authenticates a biosensor, it will send a message including the real biometric points and some chaff points. As soon as receiving the message, the biosensor will calculate the similarity between its own set of points and the set of points from the CU. If a sufficient number of points can match within a certain threshold, the biosensor will be proved legal. Then, a session key will be established. In order to improve the match accuracy, more chaff points should be transmitted, which consumes more bandwidth. At the same time, the forward secrecy and backward secrecy cannot be guaranteed. Based on our previous work, Biometric Encryption technique instead of the fuzzy vault scheme is adopted to lock the session key at one end and unlock it at the other end. As biosensors have stringent constrains of power, memory, and computation capability, high effective security technology should be implemented. When we design our key establishment protocol, security and resource constraints must be balanced.

The remainder of this paper is organized as follows. In Section 2, we describe the related work. In Section 3, we give a brief introduction to Biometric Encryption. We present the proposed scheme in Section 4. The formal verification of its correctness is presented in Section 5. The security and performance analysis are given in Section 6. Finally, we conclude the paper in Section 7.

2. Related Work

With the advance of computer and networking technology convergence trends, pervasive computing is regarded as key technology to assist real-time medical and healthcare information service with the help of deploying sensors [11, 12]. A number of theoretical and technical approaches are proposed [11, 13–21]. But only a framework approach to balance the security and the limited resources of biosensors is proposed in the above papers, lacking the real experiments on the human body signals.

Several security solutions have been proposed to protect the BAN security. Four kinds of solutions are classified in [22]: TinySec, hardware encryption, elliptic curve cryptography and biometric methods.

TinySec [23] is proposed as a solution to achieve link layer encryption in the BAN. In the TinySec approach, packets are encrypted by a group key shared among biosensors. The group key is programmed into every sensor before the sensor network is deployed. If one biosensor discloses the key or it acts as an attacker, all the information in the BAN will be disclosed.

Hardware [22] encryption is an alternative approach of TinySec. In the BAN, there is a radio chip supporting only the encryption algorithms on every biosensor. The base station shares the encryption key with every biosensor, and only the base station can decrypt the traffic. The base station collects data from the whole network and usually acts as a gateway to other networks. Therefore, it can be considered as a single point of failure. If an adversary mounts several DoS attacks, the whole WSN will be collapsed. Another drawback is that hardware encryption is highly dependent on the specific platform. Not all the biosensors produced by the different manufactures can support the platform.

Elliptic curve cryptography is a public-key cryptography approach based on the algebraic structure of elliptic curves over finite fields. It has been used in the wireless sensor network recently [24, 25]. Even though elliptic curve cryptography is feasible for sensor nodes, its energy requirements are still orders of magnitude higher compared to those of symmetric cryptosystems [22]. For example, some works [26, 27] are considered as costly due to high processing requirements. Some symmetric key distribution techniques [28, 29] require predeployment and adjust the topology when the BAN changes, which causes high computation cost because the topology of a BAN changes frequently.

Biometrics derived from the human body to secure the keying material is firstly proposed by Cherukuri in 2003 [30]. The mechanism adopts the error-correcting codes and the multiple biometrics to secure the key. Compared with the traditional asymmetric key algorithms, this technique can reduce the cost of computation and communication. In this paper, no implementation details are given. For example, how to collect the relevant biometric data and how to examine their variation with time for individuals are not expounded. The authors proposed a biometric based distributed key management approach for BAN, but only a system architecture is given without detail experiments on physiological signals [31].

Time information of heartbeats as an excellent biometric characteristic can be used to secure the BAN by Poon et al. [32, 33]. A biometric trait generated from a sequence of interpulse interval is also to secure the BAN [33]. The interpulse interval can be derived from two sources, namely, ECG and photoplethysmogram (PPG) time series. On one hand, the interpulse interval is used to secure the transmission of the encryption key between biosensors. On the other hand, it is also used as an identity for mutual authentication between biosensors. But the experiments show that the average Hamming distance between the keys generated from ECG or PPG for the same subject is 60 or 65 bits, even though the keys are long and random. The main reason is that the physiological signals from ECG and PPG have high correlation without the same values, so

the keys generated by ECG and PPG have different values. Based on the ECG signals, other schemes have been proposed for key distribution and authentication [34]. The major challenge for ECG is that biometrics derived from physiological features possesses the high degree of noise and variability inherently present in these signals. As a result, fuzzy methods are needed to enable proper operations with adequate performance in terms of false acceptance rate (FAR) and false rejection rate (FRR) [35].

The fuzzy vault scheme has so far been primarily applied to biometric authentication, such as fingerprints and iris images [36–38]. Fuzzy vault scheme is not specific to sensor networks but serves as a significant support to solve the problem of securing biosensor networks. Minhthang gave a comparative study on fuzzy vault [35]. The experiment results show that fuzzy vault scheme is suitable for securing a fuzzy key. The performance of fuzzy vault system is dependent on an error-correction code, so the specification of the error-correction code makes the design rather inflexible. Fuzzy vault is used by Agrafioti to secure the key generation between the biosensors in the BAN [7], but the computational complexity is high.

Fuzzy vault system is also used in PKSA, a scheme for authenticated pairwise key agreement between two nodes in BANs [6]. PKSA can solve the susceptivity of synchronization and feature reordering issues in [34]. The session key between two nodes is locked by using physiological signal features and fuzzy-vault cryptographic primitive and is unlocked by the receiver according to the physiological signal features measured at the other end. The key is encoded in a polynomial, which must be reconstructed based on a set of correct points to unlock the key. The polynomial degree is linear with the key length, because the key insists of the coefficients of the polynomial. If the key is too short, it is easy to guess by attackers. If the key is too long, it will take the receiver node more cost to compute the polynomial because of the presence of chaff points.

Similar to fuzzy vault, Biometric Encryption technique can be used to keep the biometrics privacy and generate a cryptographic key from biometrics [39]. Compared with fuzzy vault, the chaff points are not necessary to transmit, so the limited bandwidth in the BAN can be saved. Based on Biometric Encryption, we have designed an intradomain mutual authentication and key establishment protocol scheme [9] and an interdomain mutual authentication and key establishment protocol scheme for pervasive computing [40].

3. Preliminaries

In this section, Biometric Encryption technique will be introduced in brief.

Compared with the two classical personal authentication approaches, knowledge based approach and token based approach, Biometrics can represent the uniqueness of a user through electronic examinations of his or her physiological characteristics such as iris, fingerprint, or face and/or through behavioral characteristics. Conventional biometric identification typically consists of an enrollment stage and a verification stage. During the enrollment stage, a user's biometric template is gathered and stored in the plaintext. During the verification stage, the user's biometrics sampled on the spot is matched against the stored biometric template to verify his or her identity. If the stored biometric template of a user is compromised, there could be severe consequences for the user because the biometric template lacks revocation mechanisms.

With the proliferation of information exchange across the internet and the storage of sensitive data, cryptography has been an important technique to achieve the data confidentiality. Cryptographic algorithms are available to secure the information. Cryptographic algorithms are divided into symmetric algorithms and public-key algorithms. However, regardless of whether a symmetric or a public-key system is deployed, the system security is dependent on the key secrecy. If a key is too short, it will be guessed easily. If a key is strong enough, the large size of a key is difficult to remember. It is infeasible for a user to enter the key each time correctly. The encrypted key can be stored on a computer's hard device, but it must be protected securely.

In order to protect the users' biometric templates and keys, Mytec Technologies Inc proposed Biometric Encryption technique. Accordingly, biometric cryptosystems were originally developed for the purpose of either securing a cryptographic key using biometric features or directly generating a cryptographic key from biometric features [39]. A biometric cryptosystems also have a two-stage process: the enrollment stage and the verification stage [9, 38], as shown in Figure 1. During the enrollment stage, the biometric image is bound with a cryptographic key to create data as Bioscrypt. We refer to it as a key binding biometric cryptosystem. During the verification stage, the biometric image on the spot is combined with the Bioscrypt to recover the key. We refer to it as a key generation biometric cryptosystem.

Bioscrypt does not reveal any information about the key or biometric feature; that is, it is computationally hard to decode the key without any knowledge of the user's biometrics and vice versa. Consequently, Bioscrypt provides an excellent privacy protection. The key itself is completely independent of biometrics and can always be changed or updated. Even if the key is ever compromised, the biometrics cannot be leaked. Moreover, the key can be easily modified. In a conclusion, Biometric Encryption can not only secure a cryptographic key, but also protect the user's biometric template.

4. A Biometric Key Establishment Protocol

4.1. System Model. Our application scenario is shown in Figure 2. The notions used in describing the protocol are listed in Table 1.

Multihop communication in a BAN is most commonly used than a single hop communication in order to consume less power. The multihop structure is extremely suitable for wireless networks, especially appropriate for BAN, because each node does not require more transmitted power. Each

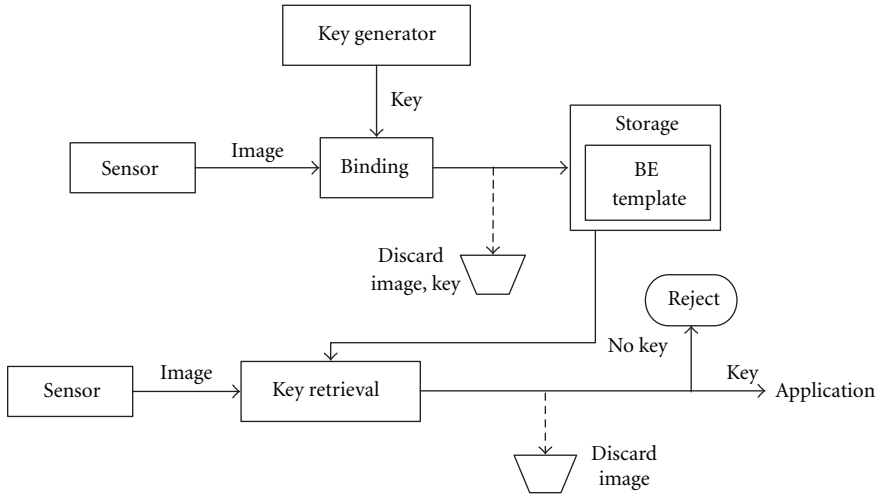


Figure 1: The process of biometric encryption.

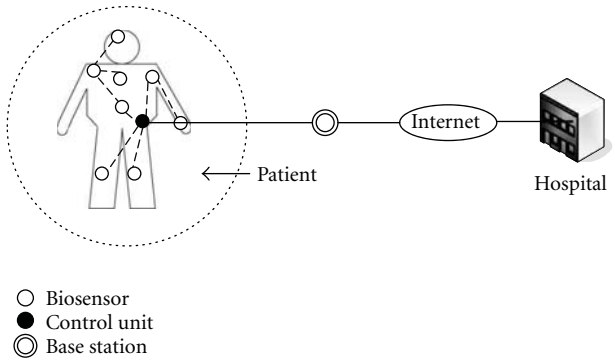


Figure 2: The system model of BANs.

Table 1: Notation table.

Symbols	Meaning
S_j	The jth biosensor
CU	Control unit
h	A secure one-way hash function
r	A random number preallocated to S_j and CU
$\{m\}_K$	A message m is encrypted by the Key K.
Na	A random number
L	The length of a random number
C_H, C_{XOR}	The computation cost of the hash function and the XOR operation
v	The order of polynomial in PSKA

node collects data and transports data to the CU via a multihop network. CU aggregates the data and sends to the remote server. All the nodes except the sink node are called biosensors. In this multihop tree, every biosensor must be authenticated by CU before transmitting the detected data. The purpose of our scheme aims to provide a mutual trust between every biosensor and CU. At the same time, a session key is generated to secure the subsequent traffic. In our experiments, ECG as the biometircs is collected and utilized. Our protocol includes two phases in Figure 3: the key binding stage and key generation stage. During key binding, a session key is bound with ECG to produce Bioscrypt. During key generation, CG can recover the session key with Bioscrypt.

4.2. Key Binding. A random number r is preallocated to CU and S_j. S_j can generate a session key to secure the subsequent traffic between CU and S_j according to the following four steps [9].

(1) S_j collects ECG signals and filters them by wavelet transform. Then, these signals are processed by discrete hashing based on the random number r [41]. Discrete hashing is described as the following.

(a) The equidistant coordinates of peak are extracted from the filtered signals, marked as (k_{xi}, k_{yi}), $i = 1, \ldots, n$, to form a feature vector $(w_i = [k_{xi} \mid k_{yi}])$. These signals are represented in a vector format, $w \in R^n$, with n denoting the feature length of w.

(b) Use r to generate m orthonormal pseudorandom vectors, $\{r_i \in R^n \mid i = 1, 2, \ldots, m\}$ and $m \leq n$.

(c) Compute the inner product $\{t^i \in T \mid i = 1, 2, \ldots, m\}$ between r and w.

(d) Compute a, m-bit code: $b^i = 0$ if $t^i \leq 0$; $b^i = 1$ if $t^i > 0$.

(2) Reed-Solomon codes are designed to correct the errors (bit differences) within the reference and test signals. Reed-Solomon codes are block-based error-correcting codes with a wide range of applications in digital communications and storage [9].

(3) Biometric template is secured by XOR process as is shown in (1). σ is Bioscrypt:

$$h(b, Na) \oplus \text{Key} = \sigma. \quad (1)$$

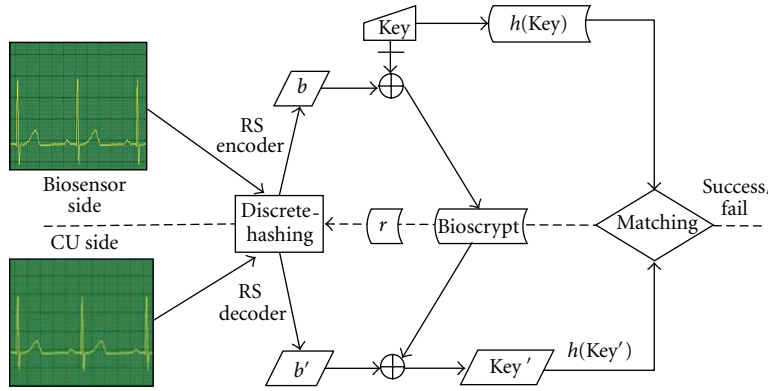


Figure 3: Biometric cryptosystem.

Fourthly, $h(\text{Key})$ as the session key is stored by the sensor while b and the filtered signals are discarded.

4.3. Key Generation. The key generation phase is comprised of the following four steps.

(1) S_j sends the following message: $S_j \rightarrow \text{CU} : \{\sigma, Na\}_{h(r)} || \{r_i\}_{h(\text{Key})}$.

(2) Because r is preallocated to CU and S_j, CU can decrypt $\{\sigma, Na\}_{h(r)}$ to get σ and Na.

(3) CU collects ECG signals filtered by using wavelet transform. These signals are also processed by discrete hashing. Reed-Solomon code is applied too. Key′ can be got in (2):

$$\sigma \oplus h(b', Na) = \text{Key}'. \tag{2}$$

(4) CU sends the following message to S_j : $\text{CU} \rightarrow S_j$: $\{r+1\}_{h(\text{Key}')}$.

(5) If $h(\text{Key})$ and $h(\text{Key}')$ are equal, S_j can decrypt $r+1$ correctly.

By these five steps, S_j and CU authenticate each other successfully and $h(\text{Key})$ is negotiated as the session key.

5. Correctness Verification

In order to ensure our protocol function correctly, the correctness of the proposed scheme is formally verified based on the SVO logic [42]. SVO logic is based on a unification of four of its logic predecessors and is relatively simple to use. SVO is a logic of belief. The intended use of SVO is to describe the beliefs of trustworthy parties involved in our protocol and the evolution of these beliefs as a consequence of communication while preserving correspondence with the original description of our protocol. In our protocol, both CU and S_j will share a fresh. In this section, the correctness of our proposed scheme is formally verified based on the SVO logic.

The symbols in SVO logic have been introduced in detail [43]. We adopt SVO to prove that both parties ascertain that they are sharing a fresh session key and both are sure that the same belief is held by the other side. Here, both CU and S_j obtain a new key, Key. Therefore, the verification goals are CU believes $S \overset{\text{Key+}}{\longleftrightarrow}$ CU and S believes $S \overset{\text{Key+}}{\longleftrightarrow}$ CU Furthermore, they believe that the new key is a fresh session key. So the verification goals are CU believes fresh(Key) and S believes fresh(Key).

First, the messages exchanged should be formalized. Then, premise sets should be figured out previously. We prove the key security from the standpoint of CU, and the process for S_j can be done right in the same way.

The Formalized Messages:

M1: $S \rightarrow \text{CU} : \{\{\sigma, Na\}_{h(r)}, \{r_i\}_{h(\text{Key})}\}$,

M2: $\text{CU} \rightarrow S : \{r_i + 1\}_{h(\text{Key}')}$.

The Premise Sets:

P1: S believes fresh(Na),

P2: S believes $S \overset{r}{\longleftrightarrow}$ CU,

P3: CU believes CU $\overset{r}{\longleftrightarrow} S$,

P4: CU believes CU has $\{r, u\}$,

P5: S believes S has $\{r, Na, \sigma, \text{Key}\}$,

P6: CU believes $PK_\delta(\text{CU}, u)$,

P7: S believes S controls $\{Na, \text{Key}\}$,

P8: CU believes (CU received $\{\{\sigma, Na\}_{h(r)}, \{r_i\}_{h(\text{Key})}\}$ $\supset$ CU received $\{\{\sigma, \text{fresh}(Na)\}_{h(r)}, \{r_i\}_{h(\text{Key})}\}$),

P9: S believes S received $\{r_i + 1\}_{h(\text{Key}')}$.

The first assumption P1 states that S_j trusts the freshness of its random number. P2 and P3 state that each principal believes their sharing random number is secure. P4 and P5 show what they have. P7 states that the principal controls the generation of the agreement key and its random number. P8 and P9 show the comprehension of CU and S_j. The verification procedures from CU's standpoint are listed as follows.

Certification Process (from CU's Standpoint):

R1: S believes $PK_\delta(S, \sigma, Na)$ // by M1, P7,

R2: CU believes CU $\overset{\text{Key}}{\longleftrightarrow}$ S // by R1, P6, Ax5, Net,

R3: CU believes (CU received $\{\sigma, \text{fresh}(Na)\}$) // by P2, P8, Ax7,

R4: CU believes CU $\overset{\text{Key}-}{\longleftrightarrow}$ S // by R2, P3, R3, R1, Ax13, Net,

R5: CU believes {S says (S sees Key)} // by M2, Ax1, Ax3, Ax17, MP,

R6: CU believes CU $\overset{\text{Key}+}{\longleftrightarrow}$ S // by R4, R5, Net,

R7: CU believes fresh(Na) // by P8, Ax1, Ax7,

R8: CU believes fresh(Key) // by R7, Ax18.

R6 and R8 show that our protocol has reached the anticipated aim. Therefore, the protocol achieves the goal of establishing a new session key for the two participants.

6. Security and Performance Analysis

In this section, we analyze the security and performance of our protocol.

6.1. Security Analysis

Data Confidentiality. The health information is so sensitive that it is important to prevent the privacy from being accessed by unauthorized entities. In our protocol, every biosensor and CU can generate a unique key to secure the traffic between them, so data confidentiality can be assured and the third party cannot decrypt it.

Data Authenticity. Data Authenticity is the property of the data by which the recipient can verify and trust that the claimed sender is a legitimate one. It is very important for BAN to prevent an illegitimate entity from masquerading as a legal one. In our protocol, the data is encrypted by h(Key) and only legal party can decrypt it.

Data Integrity. Data integrity is a property so that malicious intermediaries cannot modificate the transmission data. If a malicious entity modifies the exchange message, it will be discarded by any receiver as the message digest code cannot match due to the difference.

Mutual Authentication. The proposed scheme provides mutual trust to every bisosensor and CU. The mutual authentication is based on ECG, and a session key will be produced after authentication. The whole process has been proved by SVO.

Multiple/Cancellable/Revocable Key. Biometric Encryption can guarantee that different Bioscrypt can be got by binding the same biometrics with different keys. In our protocol, $\sigma = h(b, na) \oplus \text{Key}$ represents Bioscrypt. Even if one σ is compromised, another σ can be generated soon by binding with a new key. Biometric Encryption makes it possible to change or recompute σ easily. That is, the session key h(Key) may be revoked and replaced by newly generated one calculated from the same biometrics.

Nonlinkability. The session key h(Key) is not relevant to ECG signals and other elements. As discussed in Section 4, the session keys are independent and nonlinkable.

Forward Security. Forward Secrecy guarantees that a passive adversary who knows a contiguous subset of old session keys cannot discover subsequent session keys. In our scheme, the session key h(Key) is produced randomly by the biosensor and bound with the ECG. Thus, an adversary will have no idea of the old session keys because of the non-linkability between the session keys.

Backward Security. Backward Secrecy guarantees that a passive adversary who knows a contiguous subset of session keys cannot discover preceding the session keys. Because the session key is unlinkable, our scheme can achieve backward secrecy.

Replay Attack/Data Freshness. In order to prevent the replay attack, some time stamp or random numbers can be filled into the message to provide data freshness.

Online/Offline Guessing Attack. It is clear that a passive eavesdropper would not be able to compute the shared session key, unless he knows both the ECG signal and the random number.

The comparisons on the security features among our previous work ESKE [10], PSKA [6], and our protocol are shown in Table 2. In PSKA and ESKE, fuzzy vault scheme is adopted and the real points and fuzzy points are transferred to the receiver in plaintext. Therefore, data confidentiality, integrity, and authenticity cannot be ensured. Biometric Encryption can provide multiple, cancellable or revocable keys and non-linkability, while fuzzy vault scheme does not have this feature. Only our protocol is verified by SVO, while other works have not been proven. The table shows that our protocol can provide better security.

6.2. Performance Analysis. We evaluate our protocol using ECG physiological signal, because ECG has been found to specifically exhibit desirable characteristics for BAN applications [33]. There are existing sensor devices for medical applications, manufactured with reasonable costs, that can record these interpulse interval sequences. In our experiment, a QT database is used [44, 45], which consists of 549 holter recordings from 294 persons. These signals are sampled at 1 KHz with 16-bit resolution and over 100 fifteen-minute two-lead ECG recordings. There are onset, peak, and end markers for P, QRS, T, and U waves from 30 to 50 selected beats in each recording. The experiment platform is Matlab. In our experiments, ECG signals are processed by the wave

Table 2: Comparisons on protocol security.

Security features	PKSA [6]	ESKE [10]	Our protocol
Data confidentiality	N	N	Y
Data integrity/authenticity	N	N	Y
Mutual authentication	Y	Y	Y
Multiple/cancellable/ revocable key	N	N	Y
Nonlinkability	N	N	Y
Forward security	Y	N	Y
Backward security	Y	N	Y
Replay attack/data Freshness	Y	Y	Y
Online/offline guessing attack	Y	Y	Y
Correctness verification	N	N	Y

transform. Wavelet coefficients of different scales can be calculated by multilayer wavelet decomposition of ECG, which can remove the noise correlation coefficient and retain the useful signal components. Moreover, different rhythms have different frequencies, so multiresolution analysis separate the frequencies.

6.2.1. Distinctiveness of h(Key). An important requirement is that the physiological signals can distinguish people, which ensures that the session key h(Key) cannot be unlocked by the sensors of another person. Therefore, the physiological signals for sensors on the same subject must be clearly distinctive from those on the different subjects. FRR and FAR are used to measure the authentication accuracy. FRR is the frequency with which a genuine user is not correctly recognized and hence denied access. FAR is the frequency with which an impostor is accepted as a genuine user. If a system can accept all the ECG of right people, it will have a desirably low FRR. If a system cannot reject all the wrong people, it will have a higher FAR.

In Figure 4, we compare the FAR between our scheme and our previous work [10]. It shows that our scheme can achieve more lower FAR.

6.2.2. Robustness of h(Key). Because of the person's uniqueness, it is extremely difficult to steal and use other people ECG and it is equally difficult for an individual to mimic someone else's heart signals as they are the outcome of a combination of several sympathetic and parasympathetic factors of the human body. Thus, the session key h(Key) is robust and antiattack. When the key length is 128, 64, or 32 bits, the results in Figure 5 show that these keys are distinct for different subjects or the same subject at different time. The results also show that the session keys are nonlinkability. Even if an attacker is able to reveal one key, he cannot guess other keys because the key does not carry extra information to reveal other encryption keys. From Figure 5, we also can see that the entropy of the keys ranges from 0.586 to 1, which

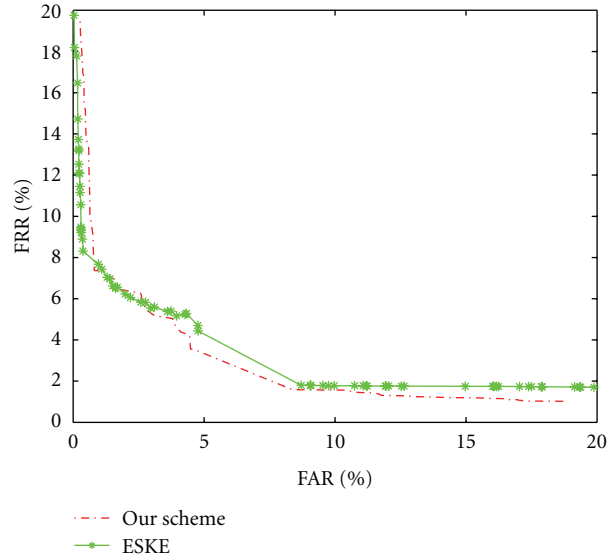


Figure 4: FAR versus FRR.

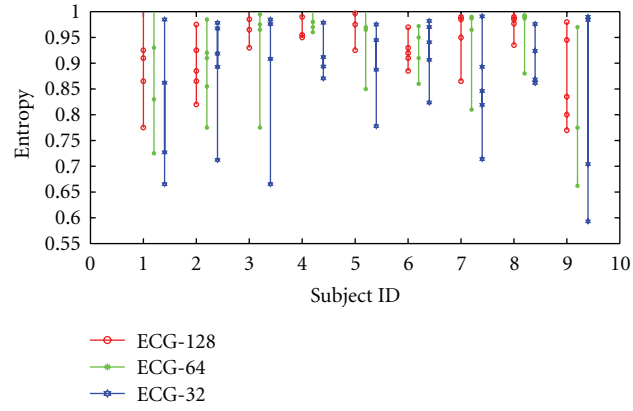


Figure 5: The entropy of different length keys.

means that our protocol possesses more uncertainty and a higher privacy level.

6.2.3. Computation and Storage Overhead. Every biosensor must store some fix and temporary parameters whose unit is byte or bit as shown in Table 3. In ESKE, a biosensor needs to store the session key and the public key of CU as fix parameters and some random numbers as temporary parameters. In PSKA, a biosensor should need the polynomial coefficients as the session key and its own identifier as the fix parameters, and some other temporary parameters. In our protocol, a biosensor stores the session key and the random r as fix parameters, w, t, b, and Na as temporary parameters.

The computation analysis in Table 4 focuses on encryptions and decryptions of asymmetric or symmetric algorithms, hash operations, XOR operation, calculated polynomial, and inner product. In ESKE, only a hash operation, a XOR operation, two public-key encryption operation in sending messages, a symmetric encryption operation to verify key, and $n + 1$ times computation of distances between

Table 3: Storage overhead.

	Fix storage	Temporary storage
ESKE [10]	L	$(m + nm + 1)L$
PSKA [6]	$(v + 1)L + L$	$(2n + 2m + 1)L$
Our protocol	$mn + L$	$n + 2m + 2L$

Table 4: Computation overhead.

	ESKE [10]	PSKA [6]	Ours protocol
Hash operation	1	1	1
XOR operation	1	0	1
Asymmetric operation	2	1	0
Symmetric operation	1	1	2
Polynomial operation	0	$m + n$	0
Inner product operation	$n + 1$	0	m

the vectors are needed. In PSKA, a hash operation to compute the message digest code, $n + m$ times computation of polynomial, and a public-key encryption operation in sending messages are needed. In our scheme, a hash operation and a XOR operation are needed to compute Bioscrypt; two symmetric operations and m times of inner product are needed.

From Tables 3 and 4, we can see that our scheme shows higher superior in storage and computation overhead compared with ESKE and PSKA. Therefore, our new scheme significantly reduces computational and storage overhead.

7. Conclusion

In this paper, we have presented a biometic key establishment scheme to protect the confidentiality and integrity of the sensitive health information. Our protocol attempts to solve the problem of security and privacy in BANs. It also aims to securely and efficiently generating and distributing the session key between a biosensor and CU.

Our protocol is based on Biometric Encryption. The primary contributions of this paper are summarized as follows.

(1) ECG is used to bind and generate a session key, which can guarantees that the key is long, random, distinctive, and temporal variant.

(2) Biometric Encryption technique is applied to derive multiple and nonlinkable keys from the same ECG signal.

(3) The correctness of the proposed scheme is formally verified based on SVO logic. Security and performance analysis show that our protocol cannot only guarantee data confidentiality, authenticity, and integrity, but also resist malicious attacks efficiently.

In the future work, we will continue to optimize our algorithm to balance safety and security.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants no. 60703101 and no. 60903153 and the Fundamental Research Funds for the Central Universities.

References

[1] E. Jovanov, A. Milenkovic, C. Otto, and P. C. de Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 2, no. 1, 2005.

[2] D. Halperin, T. S. Heydt-Benjamin, K. Fu, T. Kohno, and W. H. Maisel, "Security and privacy for implantable medical devices," *IEEE Pervasive Computing*, vol. 7, no. 1, pp. 30–39, 2008.

[3] K. Lorincz, D. J. Malan, T. R. F. Fulford-Jones et al., "Sensor networks for emergency response: challenges and opportunities," *IEEE Pervasive Computing*, vol. 3, no. 4, pp. 16–23, 2004.

[4] G. H. Zhang, C. C. Y. Poon, and Y. T. Zhang, "A biometrics based security solution for encryption and authentication in tele-healthcare systems," in *Proceedings of the 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, (ISABEL '09)*, pp. 1–4, Bratislava, Slovakia, November 2009.

[5] S. Lim, T. H. Oh, Y. B. Choi, and T. Lakshman, "Security issues on wireless body area network for remote healthcare monitoring," in *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, (SUTC '10)*, pp. 327–332, Newport Beach, Calif, USA, June 2010.

[6] K. K. Venkatasubramanian, A. Banerjee, and S. K. S. Gupta, "PSKA: usable and secure key agreement scheme for body area networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 60–68, 2010.

[7] F. Agrafioti, F. M. Bui, and D. Hatzinakos, "On supporting anonymity in a BAN biometric framework," in *Proceedings of the16th International Conference on Digital Signal Processing, (DSP '09)*, pp. 787–792, Santorini, Greece, July 2009.

[8] A. Adler, "Vulnerabilities in Biometric Encryption Systems," *Lecture Notes in Computer Science*, vol. 3546, no. 1, pp. 211–228, 2005.

[9] L. Yao, X. W. Kong, G. Wu, Q. N. Fan, and C. Lin, "A privacy-preserving authentication scheme using biometrics for pervasive computing environments," *Journal of Electronics*, vol. 27, no. 1, pp. 68–78, 2010.

[10] L. Yao, B. Liu, K. Yao, G. W. Wu, and J. Wang, "An ecg-based signal key establishment protocol in body area network," in *Proceedings of the Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, pp. 233–238, October 2010.

[11] J. Woods, "The five styles of sensory applications," Gartner Research, 2006.

[12] M. M. M. B. Amer and M. I. M. Izraiq, "System with intelligent cable-less transducers for monitoring and analyzing biosignals," *European Patent Application*, 2007.

[13] W. R. Jih, S. Y. Cheng, J. Y. J. Hsu, and T. M. Tsai, "Context-aware access control in pervasive healthcare," in *Proceedings of the EEE'05 Workshop: Mobility, Agents, and Mobile Services (MAM)*, Hong Kong, China, March 2005.

[14] K. Adam, B. Price, M. Richards, and B. Nuseibeh, "A privacy preference model for pervasive computing," in *Proceedings of the Euro mGov 2005*, Brighton, UK, July 2005.

[15] G. Zhang and M. Parashar, "Context-aware dynamic access control for pervasive applications," in *Proceedings of the Communication Networks and Distributed Systems Modeling and Simulation Conference, (CNDS '04)*, Western MultiConference (WMC), San Diego, Calif, USA, January 2004.

[16] D. Jea, I. S. Yap, and M. B. Srivastava, "Context-aware access to public shared devices," in *Proceedings of the HealthNet 2007: The First International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments*, Puerto Rico, USA, June 2007.

[17] M. Tentori, J. Favela, and M. D. Rodriguez, "Privacy-aware autonomous agents for pervasive healthcare," *IEEE Intelligent Systems*, vol. 21, no. 6, pp. 55–62, 2006.

[18] M. Haque and S. I. Ahamed, "Security in pervasive computing: current status and open issues," *International Journal of Network Security*, vol. 3, no. 3, pp. 203–214, 2006.

[19] S. C. Shin, C. Y. Ryu, J. H. Kang et al., "Realization of an e-health system to perceive emergency situations," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (EMBC '04)*, vol. 2, pp. 3309–3312, San Francisco, Calif, USA, September 2004.

[20] R. Bose, A. Helal, S. Lim, and V. S. Sivakumar, "Virtual sensors for service oriented intelligent environments," in *Proceedings of the 3rd IASTED International Conference on Advances in Computer Science and Technology, (ACST '07)*, Phuket, Thailand, April 2007.

[21] S. Lim, S. Kang, and J. Sohn, "Modeling multiple agent based cryptographic key recovery protocol," in *Proceedings of the 19th Annual Computer Security Applications Conference, (ACSAC '03)*, pp. 119–128, Las Vegas, Nev, USA, December 2003.

[22] M. Mana, M. Feham, and B. A. Bensaber, "Trust key management scheme for wireless body area networks," *International Journal of Network Security*, vol. 12, no. 2, pp. 61–69, 2011.

[23] C. Karlof, N. Sastry, and D. Wagner, "TinySec: a link layer security architecture for wireless sensor networks," in *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems, (SenSys '04)*, pp. 162–175, Baltimore, Md, USA, November 2004.

[24] M. Guennoun, M. Zandi, and K. El-Khatib, "On the use of biometrics to secure wireless biosensor networks," in *Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications, (ICTTA '08)*, pp. 1–5, Damascus, Syria, April 2008.

[25] P. Szczechowiak, L. B. Oliveira, M. Scott, M. Collier, and R. Dahab, "NanoECC: testing the limits of elliptic curve cryptography in sensor networks," in *Proceedings of the 5th European Conference on Wireless Sensor Networks*, pp. 305–320, Bologna, Italy, February 2008.

[26] Q. Huang, J. Cukier, H. Kobayashi, B. Liu, and J. Zhang, "Fast authenticated key establishment protocols for self-organizing sensor networks," in *Proceedings of the International Workshop on Wireless Sensor Networks and Applications, (WSNA '03)*, pp. 141–150, San Diego, Calif, USA, September 2003.

[27] D. J. Malan, M. Welsh, and M. D. Smith, "A public-key infrastructure for key distribution in tinyos based on elliptic curve cryptography," in *Proceedings of the First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, (SECON '04)*, Santa Clara, Calif USA, October 2004.

[28] F. Adelstein, S. K. S. Gupta, G. G. Richard, and L. Schwiebert, *Fundamentals of Mobile and Pervasive Computing*, McGraw-Hill, New York, NY, USA, 2005.

[29] S. Zhu, S. Setia, and S. Jajodia, "LEAP: efficient security mechanisms for large-scale distributed sensor networks," in *Proceedings of the 10th ACM Conference on Computer and Communications Security, (CCS '03)*, vol. 2, pp. 500–528, Washington, DC, USA, October 2003.

[30] S. Cherukuri, K. K. Venkatasubramanian, and S. K. S. Gupta, "BioSec: a biometric based approach for securing communication in wireless networks of biosensors implanted in the human body," in *Proceedings of the International Conference on Parallel Processing Workshops*, pp. 432–439, Kaohsiung, Taiwan, October 2003.

[31] S. M. K.-U.-R. Raazi, H. Lee, S. Lee, and Y.-K. Lee, "Bari+: a biometric based distributed key management approach for wireless body area networks," *Sensors*, vol. 10, no. 4, pp. 3911–3933, 2010.

[32] C. C. Y. Poon, Y.-T. Zhang, and S.-D. Bao, "A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 73–81, 2006.

[33] S. D. Bao, C. C. Y. Poon, L. F. Shen, and Y. T. Zhang, "Using the timing information of heartbeats as an entity identifier to secure body sensor network," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 772–779, 2008.

[34] F. M. Bui and D. Hatzinakos, "Biometric methods for secure communications in body sensor networks: resource-efficient key management and signal-level data scrambling," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 529879, 16 pages, 2008.

[35] F. M. Bui and D. Hatzinakos, "Secure methods for fuzzy key binding in biometric authentication applications," in *Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers, (ASILOMAR '08)*, pp. 1363–1367, Pacific Grove, Calif, USA, October 2008.

[36] U. Uludag, S. Pankanti, and A. K. Jain, "Fuzzy vault for fingerprints," in *Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA '05)*, vol. 3546, pp. 310–319, Hilton Rye Town, NY, USA, July 2005.

[37] E. S. Reddy and I. R. Babu, "Authentication using fuzzy vault based on iris textures," in *Proceedings of the 2nd Asia International Conference on Modelling and Simulation, (AMS '08)*, pp. 361–368, Kuala Lumpur, Malaysia, May 2008.

[38] A. Juels and M. Sudan, "A fuzzy vault scheme," in *Proceedings of the International Symposium on Information Theory*, vol. 38, pp. 237–257, Seattle, Wash, USA, July 2006.

[39] A. K. Jain, K. Nandakumar, and A. Nagar, "Biometric template security," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 579416, 17 pages, 2008.

[40] L. Yao, L. Wang, X. W. Kong, G. W. Wu, and F. Xia, "An inter-domain authentication scheme for pervasive computing environment," *Computers and Mathematics with Applications*, vol. 60, no. 2, pp. 234–244, 2010.

[41] A. T. B. Jin, D. C. L. Ngo, and A. Goh, "Biohashing: two factor authentication featuring fingerprint data and tokenised random number," *Pattern Recognition*, vol. 37, no. 11, pp. 2245–2255, 2004.

[42] C. Meadows, "Formal methods for cryptographic protocol analysis: emerging issues and trends," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 1, pp. 44–54, 2003.

[43] P. Syverson and I. Cervesato, "The logic of authentication protocols," *Foundations of Security Analysis and Design*, vol. 2171, pp. 63–136, 2001.

[44] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "Database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," *Computers in Cardiology*, vol. 27, pp. 673–676, 1997.

[45] A. L. Goldberger, L. A. N. Amaral, L. Glass et al., "PhysioBank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, 2000.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 828414, 19 pages
doi:10.1155/2011/828414

Research Article

Anonymous Aggregator Election and Data Aggregation in Wireless Sensor Networks

Tamás Holczer and Levente Buttyán

Laboratory of Cryptography and Systems Security (CRYSYS), Budapest University of Technology and Economics, Bme-Hit, P.O. Box 91, 1521 Budapest, Hungary

Correspondence should be addressed to Tamás Holczer, holczer@crysys.hu

Received 15 February 2011; Revised 15 May 2011; Accepted 30 June 2011

In mission critical cyber-physical systems, dependability is an important requirement at all layers of the system architecture. In this paper, we propose protocols that increase the dependability of wireless sensor networks, which are potentially useful building blocks in cyber-physical systems. More specifically, we propose two private aggregator node election protocols, a private data aggregation protocol, and a corresponding private query protocol for sensor networks that allow for secure in-network data aggregation by making it difficult for an adversary to identify and then physically disable the designated aggregator nodes. Our advanced protocols resist strong adversaries that can physically compromise some nodes.

1. Introduction

Wireless sensor and actuator networks are potentially useful building blocks for cyber-physical systems. Those systems must typically guarantee high-confidence operation, which induces strong requirements on the dependability of their building blocks, including the wireless sensor and actuator network. Dependability means resistance against both accidental failures and intentional attacks, and it should be addressed at all layers of the network architecture, including the networking protocols and the distributed services built on top of them, as well as the hardware and software architecture of the sensor and actuator nodes themselves. Within this context, in this paper, we focus on the security aspects of aggregator node election and data aggregation protocols in wireless sensor networks.

Data aggregation in wireless sensor networks helps to improve the energy efficiency and the scalability of the network. It is typically combined with some form of clustering. A common scenario is that sensor readings are first collected in each cluster by a designated aggregator node that aggregates the collected data and sends only the result of the aggregation to the base station. In another scenario, the base station may not be present permanently in the network, and the aggregated data must be stored by the designated aggregator node in each cluster temporarily until the base station can eventually fetch the data. In both cases, the amount of communication, and hence, the energy consumption of the network can be greatly reduced by sending aggregated data, instead of individual sensor readings, to the base station.

While data aggregation in wireless sensor networks is clearly advantageous with respect to scalability and efficiency, it introduces some security issues. In particular, the designated aggregator nodes that collect and store aggregated sensor readings and communicate with the base station are attractive targets of physical node destruction and jamming attacks. Indeed, it is a good strategy for an attacker to locate those designated nodes and disable them, because he can prevent the reception of data from the entire cluster served by the disabled node. Even if the aggregator role is changed periodically by some election process, some security issues remain, in particular in the case when the base station is off-line and the aggregator nodes must store the aggregated data temporarily until the base station goes on-line and retrieves them. More specifically, in this case, the attacker can locate and attack the node that was aggregator in a specific time epoch before the base station fetches its stored data, leading to permanent loss of data from the given cluster in the given epoch.

In order to mitigate this problem, we introduced the concept of private aggregator node election, and we proposed the first private aggregator node election protocol in our earlier work [1]. Briefly, our earlier protocol ensures that the identity of the elected aggregator remains hidden from an attacker who observes the execution of the election process. However, our earlier protocol ensures only protection against an external eavesdropper that cannot compromise sensor nodes, and it does not address the problem of identifying the aggregator nodes by means of traffic pattern analysis after the election phase.

In our second paper [2], we addressed the shortcomings of our earlier scheme: we proposed a new private aggregator node election protocol that is resistant even to internal attacks originating from compromised nodes, and we also proposed a new private data aggregation protocol and a new private query protocol which preserved the anonymity of the aggregator nodes during the data aggregation process and when they provide responses to queries of the base station. In our second private aggregator node election protocol, each node decides locally in a probabilistic manner to become an aggregator or not, and then the nodes execute an anonymous veto protocol to verify if at least one node became aggregator. The anonymous veto protocol ensures that nonaggregator nodes learn only that there exists at least one aggregator in the cluster, but they do not learn any information on its identity. Hence, even if such a nonaggregator node is compromised, the attacker learns no useful information regarding the identity of the aggregator.

However, our second protocol used a special broadcast communication scheme, which assumed the existence of a Hamilton cycle in the cluster. Creating a Hamilton cycle in a wireless sensor network is a difficult and energy-consuming problem, therefore in this paper, we relax this assumption and we use spanning trees, which are easy to construct.

Another important improvement in this paper is that we address the problem of misbehaving nodes that try to obfuscate the values received by the operator. We analyze how a malicious node can mislead the operator and propose an algorithm which can detect the presence of misbehaving nodes.

Our protocols can be used to protect sensor network applications that rely on data aggregation in clusters, and where locating and then disabling the designated aggregator nodes is highly undesirable. Such applications include high-confidence cyber-physical systems where sensors and actuators monitor and control the operation of some critical physical infrastructure, such as an energy distribution network, a drinking water supply system, or a chemical pipeline. A common feature of these systems is that they have a large geographical span, and therefore, the sensor network must be organized into clusters and use in-network data aggregation in order to ensure scalability and energy-efficient operation. Moreover, due to the mission critical nature of these applications, it is desirable to prevent the identification of the aggregator nodes in order to limit the impact of a successful attack against the sensor network. Our first protocol that resists only an external eavesdropper is less complex than our second protocol that works in a stronger attacker model. Hence, the first protocol can be used in case of strong resource constraints or when the risk of compromising sensor nodes is limited (e.g., it may be difficult to obtain physical access to the nodes). Our second protocol is needed when the risk of compromised and misbehaving nodes cannot be eliminated by other means.

The remainder of the paper is organized as follows. In Section 2, we introduce our system and attacker models. In Section 3, we present our basic aggregator election protocol which can withstand external attacks, while in Section 4, we introduce our advanced protocols, which can withstand internal aggregator identification and scamming attackers as well. In Section 5, we give an overview of some related work, and in Section 6, we conclude the paper and sketch some future research directions.

2. System and Attacker Models

A sensor network consists of sensor nodes that communicate with each other via wireless channels. Every node can generate sensor readings and store it or forward it to another node. Each node can directly communicate with the nodes within its radio range; those nodes are called the (one-hop) neighbors of the node. In order to communicate with distant nodes (outside the radio range), the nodes use multi-hop communications. The sensor network has an operator as well, who can communicate with some of the nodes through a special node called base station, or can communicate directly with the nodes if the operator moves close to the network.

Throughout the paper, a data-driven sensor network is envisioned, where every sensor node sends its measurement to a data aggregator regularly. Such data driven networks are used for regular inspection of monitored processes notably in critical infrastructures. Event-driven networks can be used for reporting special usually dangerous but infrequent events like fire in a building. There is no need of clustering and data aggregation in event-based systems, thus private cluster aggregator election and data aggregation is not applicable there. The third kind of network is the query-driven network, where the operator sends a query to the network, and the network sends a response. This kind of functionality can be used with data-driven networks, and can have privacy consequences, like the identity of the answering node should remain hidden.

In the following, we will assume, that the time is slotted, and one measurement is sent to the data aggregator in each time slot. The time synchronization between the nodes is not discussed here, but a comprehensive survey can be found in [3].

It is assumed that every node shares some cryptographic credentials with the operator. These credentials are unique for every node, and the operator can store them in a lookup table, or can be generated from a master key and the node's identifier on demand. The exact definition of the credentials can be found in Sections 3.1 and 4.1.

The nodes may be aware of their geographical locations, and they may already be partitioned into well-defined

geographical regions. In this case, these regions are the clusters, and the objective of the aggregator election protocol is to elect an aggregator within each geographical region. We call this approach location-based clustering; an example would be the PANEL protocol [4].

A kind of generalization of the position-based election is the preset case, where the nodes know the cluster ID they belong to before any communication. Here the goal of the election is to elect one node in every preset cluster. This approach is used in [2].

Alternatively, the nodes may be unaware of their locations or cluster IDs and know only their neighbors. In this case, the clusters are not predetermined, but they are dynamically constructed parallel to the election of the aggregators. Basically, any node may announce itself as an aggregator, and the nodes within a certain number of hops on the topology graph may join that node as cluster members. We call this approach topology-based clustering; an example would be the LEACH protocol [5].

The location-based and the topology-based approaches are illustrated in Figure 1.

Both approaches may use controlled flooding of broadcast messages. In case of location-based or preset clustering, the scope of a flood is restricted to a given geographic region or preset cluster. Nodes within that region rebroadcast the message to be flooded when they receive it for the first time. Nodes outside of the region or having different preset cluster IDs simply drop the message. In case of topology-based clustering, we assume that the broadcast messages has a time-to-live field that controls the scope of the flooding. Any node that receives a broadcast message with a positive TTL value for the first time will automatically decrement the TTL value and rebroadcast the message. Duplicates and messages with TTL smaller than or equal to zero are silently discarded. When we say that a node broadcasts a message, we mean such a controlled flooding (either location-based, preset, or topology-based, depending on the context). In Section 4, we will use connected dominating sets (CDSs) to implement efficient broadcast messaging. The concept of CDS will be introduced there.

We call the set of nodes which are (in the location-based and the preset case) or can potentially be (in the topology-based case) in the same cluster as a node S the *cluster peers* of S. Hence, in the location-based case, the cluster peers of S are the nodes that reside within the same geographic region as node S. In the preset case, the cluster peers are the nodes sharing the same cluster ID. In the topology-based case, the set of cluster peers of S usually consists in its n-hop neighborhood, for some parameter n. The nodes may not explicitly know all their cluster peers.

The main functional requirement of any clustering algorithm is that either node S or at least one of the cluster peers of S will be elected as aggregator.

The leader of each cluster is called cluster aggregator, or simply aggregator. In the following we will use aggregator, cluster aggregator, and data aggregator interchangeably.

As mentioned in Section 1, an attacker can gain much more information by attacking an aggregator node than attacking a normal node. To attack a data aggregator node either physically or logically, first the attacker must identify that node. In this paper we assume that the attacker's goal is to identify the aggregator (which means that simply preventing, jamming, or confusing the aggregation is not the goal of the attacker). In Section 4.5 we go a little further, and analyze what happens if a compromised node does not follow the proposed protocols in order to mislead the operator.

An attacker who wants to discover the identity of the aggregators can eavesdrop the communication between any nodes, can actively participate in the communication (by deleting modifying and inserting messages) and can physically compromise some of the nodes. A compromised node is under the full control of the attacker, the attacker can fully review the inner state of that node, and can control the messages sent by that node.

Compromising a node is a much harder challenge for an attacker then simply eavesdropping the communication. It requires physical contact with the node and some advanced knowledge, however it is far from impossible for an attacker with good electrical and laboratory background [6]. So we propose two solutions. The first basic protocol can fully withstand a passive eavesdropper, but a compromising attacker can gain some knowledge about the identities of the cluster aggregators. The second advanced protocol can withstand a compromising attacker as well, with only leaking information about the compromised nodes.

In case of a passive adversary, a rather simple solution could be based on a commonly shared global key. Using that shared global key as a seed of a pseudorandom number generator, every node can construct locally (without any communications) the same pseudo randomly ordered list of all nodes. These lists will be identical for every node because all nodes use the same seed and the same pseudo random number generator. Then, the first A nodes of the list are elected aggregators such that every node can communicate with a cluster aggregator and no subset of A covers the whole system. An illustration of the result of this algorithm can be seen on Figure 1 for location-based and topology-based cluster aggregator election.

The problem with this solution is that it is not robust: compromising a single node would leak the common key, and the adversary could compute the identifier of all cluster aggregators. While we do not want to fully address the problem of compromised nodes in the first protocol, we still aim at a more robust solution than the one described above. In particular, the system should not collapse by compromising just a single or a few nodes.

The second protocol can withstand the compromise of some nodes without the degradation of the privacy of the cluster aggregators. This protocol meets the following goals and has the following limitations.

(i) The identity of the noncompromised cluster aggregators remains secret even in the presence of passive and active attackers or compromised nodes.

(ii) The attacker can learn whether the compromised node is an aggregator.

(iii) An attacker can force a compromised node to be aggregator, but does not know anything about the existence or identity of the other aggregators.

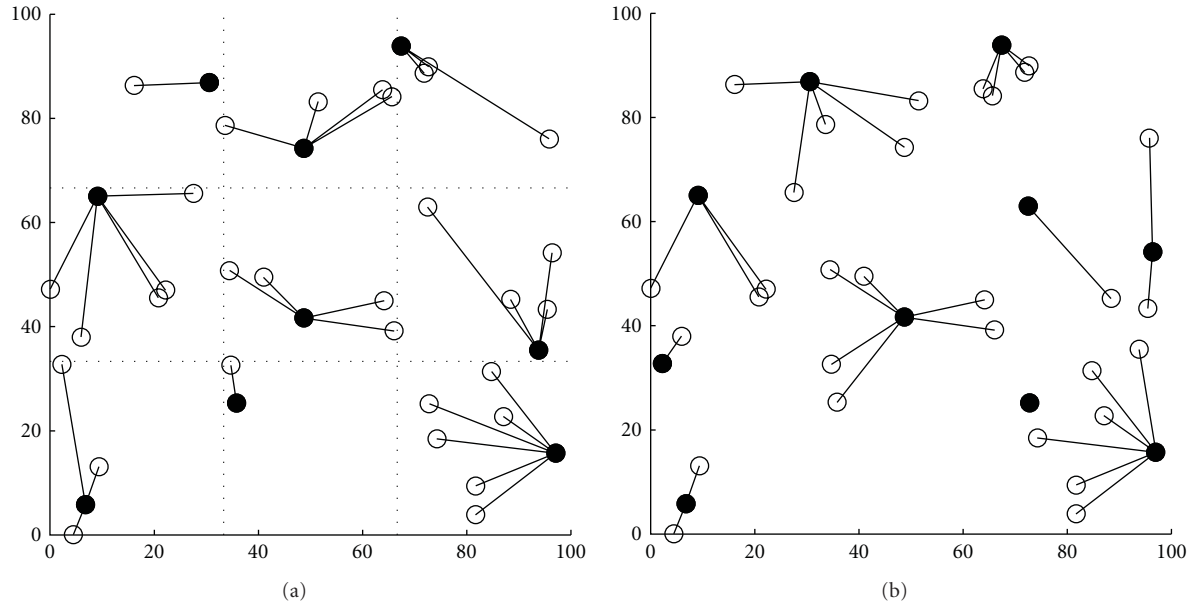


Figure 1: Result of a location-based (a) and topology-based (b) one-hop aggregator election protocol. Solid dots represent the aggregators, and empty circles represent cluster members.

(iv) The attacker cannot achieve that no aggregator is elected in the cluster, however all the elected aggregator(s) may be compromised nodes.

The main difference between the first and second protocol is the following. The first protocol is very simple, but not perfect as a compromised node can reveal the identity of the aggregators. The second protocol requires more complex computations, but offers anonymity in case of node compromise as well. In some cases such complex computations are outside the capabilities of the nodes (or the probability of compromise is low), but anonymity is still required by the system. In these cases we suggest to use the first protocol. If the probability of node compromise is not negligible, then the use of the second protocol is recommended.

3. Basic Protocol

In this section, we describe the basic protocol that we propose for private aggregator node election. An important component of this basic protocol will be presented in Section 3.2, where we also describe how to set the parameters of the protocol. We first give a brief overview of the basic principles of our protocol and present the details later.

3.1. Protocol Description. We assume that the nodes are synchronized (see [3] for a survey on time synchronization mechanism for sensor networks), and each node starts executing the protocol roughly at the same time. The protocol terminates after a predefined fix amount of time. During the execution of the protocol, any node that has not received any aggregator announcement yet may decide to become an aggregator, in which case, it broadcasts an aggregator announcement message announcing itself as a cluster aggregator. This message is broadcast among the cluster peers of the node sending the announcement (see Section 2). Upon reception of a cluster aggregator announcement, any node that has neither announced itself as a cluster aggregator nor received any such announcement yet will consider the sender of the announcement as its cluster aggregator. In order to prevent an external observer to learn the identity of the cluster aggregators, all messages sent in the protocol are encrypted such that only the nodes to whom they are intended can decrypt them. For this, we assume that each node shares a common key with all of its cluster peers (an overview of available key establishment mechanisms for sensor networks can be found in [7]). In addition, in order to avoid that message originators are identified as cluster aggregators, the nodes that will be cluster members are required to send dummy messages that cannot be distinguished from the announcements by the external observer (i.e., they are encrypted and disseminated in the same way as the announcements).

Note that the proposed basic protocol considers only either pairwise keys between the neighboring nodes or group keys shared between sets of neighboring nodes, so no global key is assumed. Such pairwise or group keys can be established by the techniques proposed in [7]. The key establishment can be based on randomly selected key sets. In such a protocol, the probability that neighboring nodes share a common key is high, and the unused keys are deleted [8]. The key establishment can be also based on a common key which is deleted after some short time when the neighbors are discovered [9]. Any node that owns the common key can generate a pairwise key with a node which owns or previously owned the common key. The basic method for exchanging a group/cluster key with the neighboring nodes is to send

```
start T1, expires in rand(0,τ) //timer, expires in round 1
start T2, expires in rand(τ,2τ) //timer, expires in round 2
announFirst = (rand(0,1) ≤ γ)
CAID = −1 // ID of the cluster aggregator of the node
while T1 NOT expired do
   if receive ENC(announcement) AND (CAID = −1) then
      CAID = ID of sender of announcement
   end if
end while
// T1 expired
if announFirst AND (CAID = −1) then
   broadcast ENC(announcement);
   CAID = ID of node itself;
else
   broadcast ENC(dummy);
end if
while T2 NOT expired do
   if receive ENC(announcement) AND (CAID = −1) then
      CAID = ID of sender of announcement
   end if
end while
// T2 expired
if (NOT announFirst) AND (CAID = −1) then
   broadcast ENC(announcement);
   CAID = ID of node itself;
else
   broadcast ENC(dummy);
end if
```

Algorithm 1: Basic private cluster aggregator election algorithm.

the same random key to each neighbor encrypted with the previously exchanged pairwise keys.

The pseudocode of the protocol is given in Algorithm 1, and a more detailed explanation of the protocol's operation is presented below. The protocol consists of two rounds, where the length of each round is τ. The nodes are synchronized; they all know when the first round begins, and what the value of τ is. At the beginning, each node starts two random timers, T1 and T2, where T1 expires in the first round (uniformly at random) and T2 expires in the second round (uniformly at random). Each node also initializes at random a binary variable, called announFirst, that determines in which round the node would like to send a cluster aggregator announcement. The probability that announFirst is set to the first round is γ, which is a system parameter. The setting of γ is elaborated in Section 3.2.

In the first round, every node S waits for its first timer T1 to expire. If S receives an announcement before T1 expires, then the sender of the announcement will be the cluster aggregator of S. When T1 expires, S broadcasts a message as follows: if announFirst is set to the first round and S has not received any announcement yet, then S sends an announcement, in which it announces itself as a cluster aggregator. Otherwise, S sends a dummy message. In both cases, the message is encrypted (denoted by ENC() in the algorithm) such that only the cluster peers of S can decrypt it.

Table 1: Estimated time of the building blocks on a MICAz.

Algorithm	Generation [ms]	Verification [ms]
SHA-1 [10]	1.4	—
RSA 1024 bit [11]	12040	470
RC4 [10]	0.1	0.1
RC5 [10]	0.4	0.4

The second round is similar to the first round. When T2 expires S broadcasts a message as follows: If announFirst is set to the second round and S has not received any announcement yet, then S sends an announcement, otherwise, S sends a dummy message. In both cases, the message is encrypted.

It is easy to see that at the end of the second round each node is either a cluster aggregator or it is associated with a cluster aggregator whose ID is stored in variable CAID. Without the second round, a node can remain unassociated, if it sends and receives only dummy messages in the first round. In addition, a passive observer only sees that every node sends two encrypted messages, one in each round. This makes it difficult for the adversary to identify who the cluster aggregators are (see also more discussion on this in the next section). In addition, if a node is compromised, the adversary learns only the identity of the cluster aggregators whose announcements have been received by the compromised node.

In WSNs, it must be analyzed what happens if some messages are delayed or lost in the noisy unreliable channel. Two cases must be analyzed, dummy messages and announcements. If a dummy message is delayed or not delivered successfully to all recipients, then the result of the protocol is not modified as dummy messages serve for only covering the announcements. If an announcement is delayed or not delivered to a node, then the recipient will not select the sender as cluster aggregator. It will select a node who sent the announcement later or the node elects itself and sends an announcement. The message loss may modify the resulting set of cluster aggregators, but neither harm the anonymity of the elected aggregators, nor harm the original goal of cluster aggregator election (a node must be either a cluster aggregator or a cluster aggregator must be elected from the nodes cluster peers).

Note that two neighboring nodes can send an announcement at the same time with some small probability. Actually, it is not a problem in the protocol. The only result is that both nodes will be cluster aggregators independently. As it is not conflicting with the original goal of cluster aggregator election, this infrequent situation does not need any special attention.

The overhead introduced by the basic protocol is sending two encrypted messages for each election round. Other protocols [4, 5] use one (or zero) unencrypted messages to elect an aggregator. So the number of messages sent in the election phase is slightly larger compared to other solutions. The symmetric encryption also causes some extra overhead (for details, see Table 1, rows with RC4 and RC5).

3.2. Protocol Analysis. In this section, the previously suggested basic protocol is analyzed. As defined in Section 2, the main goal of the attacker is to reveal the identity of the cluster aggregators. To do so, the attacker can eavesdrop modify and delete messages, and can capture some nodes.

First the logical attacks are analyzed where the attacker does not capture any nodes, then the results of a node capture.

The attackers main goal is to reveal the identity of the cluster aggregators. As all the internode communication is encrypted and authenticated, it cannot get any information from the messages themselves, but it can get some side information from simple traffic and topology analysis.

3.2.1. Density-Based Attack. Thanks to the dummy messages and the encryption in the basic protocol, an external observer cannot trivially identify the cluster aggregators; however, it can still use side information and suspect some nodes to be cluster aggregators with higher probability than some other nodes. Such a side information is the number of the cluster peers of the nodes. This number correlates with the local density of the nodes, that is why this attack is called density-based attack. Indeed, the probability of becoming a cluster aggregator depends on the number of the cluster peers of the node. For instance, if a node does not have any cluster peers, it will be a cluster aggregator with probability one. On the other hand, if the node has a larger number of cluster peers, then the probability of receiving an announcement from a cluster peer is large, and hence, the probability that the node itself becomes cluster aggregator is small. Note also that the number of cluster peers can be deduced from the topology of the network, which may be known to the adversary.

The probability of becoming a cluster aggregator is approximately inversely proportional to the number of cluster peers:

$$\Pr(\mathrm{CA}(S)) \cong \frac{1}{D(S)}, \tag{1}$$

where $\mathrm{CA}(S)$ is the event of S being elected cluster aggregator, and $D(S)$ is the number of cluster peers of node S. Figure 2 illustrates this proportionality where the curve belongs to (1) and the plotted dots correspond to simulation results (100 nodes, random deployment, one hop communication, and topology-based clustering).

Two approaches can be used to mitigate this problem. One is to take the number of cluster peers of the nodes into account when generating the random timers for the protocol. The second is to balance the logical network topology in such a way that every node has the same number of cluster peers. In the following a possible solution for both approaches is introduced.

The first approach can be the fine tuning of the distributions. It is not analyzed here deeply. It can only slightly modify the probabilities of being cluster aggregator, so it has no large effects. An example can be seen on Figure 3, where the 10th power of $D(S)$ is used as a normalizing factor, when γ (probability of sending an announcement in the first round) is computed. The coefficients of the polynomial are

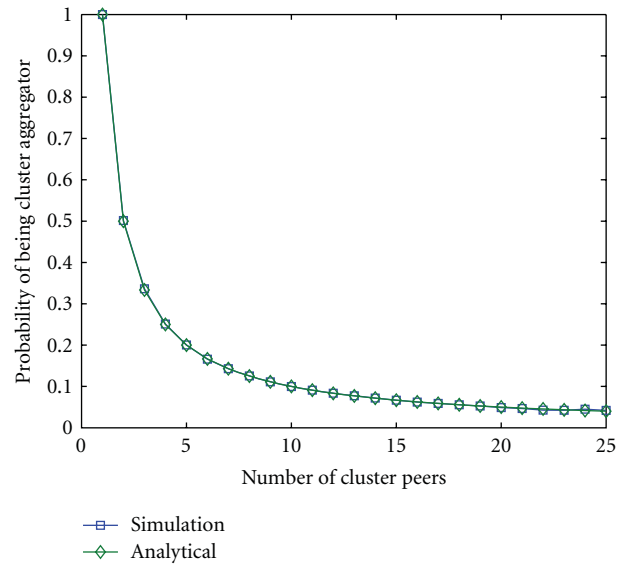


Figure 2: Probability of being cluster aggregator as a function of the number of cluster peers.

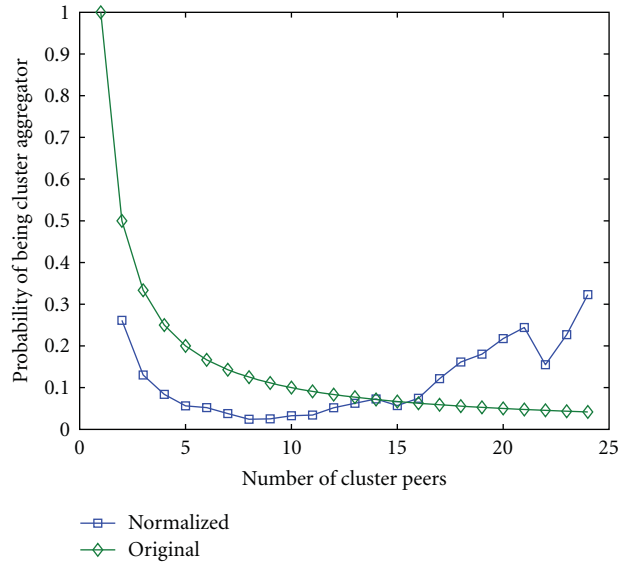


Figure 3: Probability of being cluster aggregator as a function of number of cluster peers. The analytical values comes from (1), while the simulation values come from simulation, where the γ probabilities are normalized with the number of cluster peers of the nodes.

set as resulting curve is the closest to uniform distribution. It can be seen that modifying γ on a per node basis does not eventually reach its goal; the normalized distribution is far from uniform. Actually by modifying γ, the other attack discussed in the next section can be mitigated, so here we propose a solution which does not set the γ parameter.

The second approach modifies the number of cluster peers of a node to reach a common value. Let us denote this value by α.

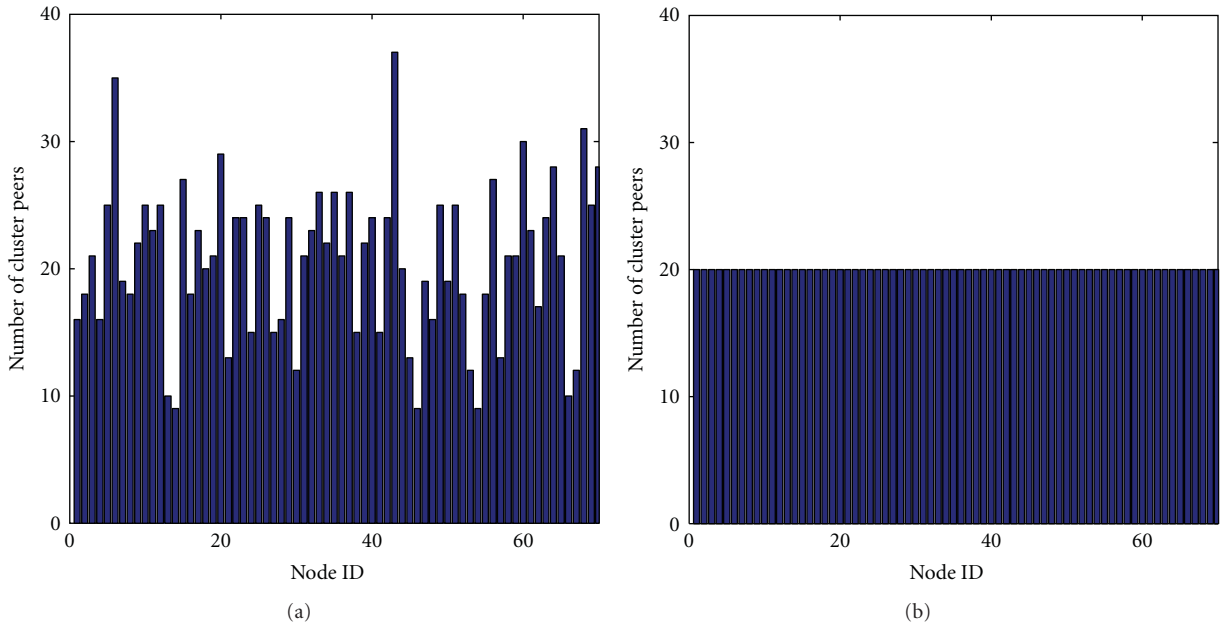


Figure 4: Result of balancing. The 70 nodes are represented on the x-axis. The number of cluster peers before (a), and after (b) the balancing are represented on the y-axis.

An efficient approach to mitigate this problem is to modify the number of cluster peers such that it becomes a common value α for all of them. In theory, this common value can be anything between 1 and the total number N of the nodes in the network. In practice, it should be around the average number of cluster peers, which can be estimated locally by the nodes. For example, assuming one-hop communications (meaning that the cluster peers are the radio neighbors), the following formula can be used:

$$\alpha = (N-1)\frac{R^2\pi}{A} + 1 \simeq E(D(S)), \tag{2}$$

where R is the radio range, and A is the size of the total area of the network. The formula is based on the fact that the number of cluster peers is proportional to the ratio between radio coverage and total area. Similar formulae can be derived for the general case of multi-hop communication.

If a node S has more than α cluster peers it can simply discard the messages from $D(S) - \alpha$ randomly chosen cluster peers. If S has less than α cluster peers it must get new cluster peers by the help of its actual cluster peers (if S has not got any cluster peers originally, then it will always become a cluster aggregator). The new cluster peers can be selected from the set of cluster peers of the original cluster peers. To explore the potential new cluster peers, every node can broadcast its list of cluster peers within its few hop neighborhood before running the basic protocol. From the lists of the received cluster peers, every node can select its $\alpha - D(S)$ new cluster peers uniformly at random. Then, the basic aggregator election protocol can be executed using the balanced set of cluster peers. An example for this balancing is shown in Figure 4 (70 nodes, random deployment, one hop communication, and topology-based clustering).

After running the balancing protocol, every node can approach the envisioned α value. The advantage of the balancing protocol is that however an attacker can gather the information about the number of cluster peers, this number is efficiently balanced after the protocol. The drawback of this solution is that it requires the original cluster peers to relay messages between distant nodes. One can imagine this solution as selectively increasing the TTL of protocol messages creating much larger neighborhoods.

3.2.2. Order-Based Attack. Another important side information an attacker can use is the *order* in which the nodes send messages in the first round of the protocol. Indeed, the sender of the ith message will be cluster aggregator if none of the previous $i-1$ messages are announcements (but dummies) and the ith message is an announcement. Thus, the probability P_i that the sender of the ith message becomes cluster aggregator depends on i and parameter γ:

$$P_i = (1-\gamma)^{i-1}\gamma, \quad 1 \le i \le n. \tag{3}$$

The $(n+1)$th element of the distribution is the probability that no announcement is sent in the first round:

$$P_{n+1} = (1-\gamma)^n \tag{4}$$

in which case the sender of the first message of the second round must be a cluster aggregator.

The entropy of this distribution characterizes the uncertainty of the attacker who wants to identify the cluster aggregator using the order information. Assuming that

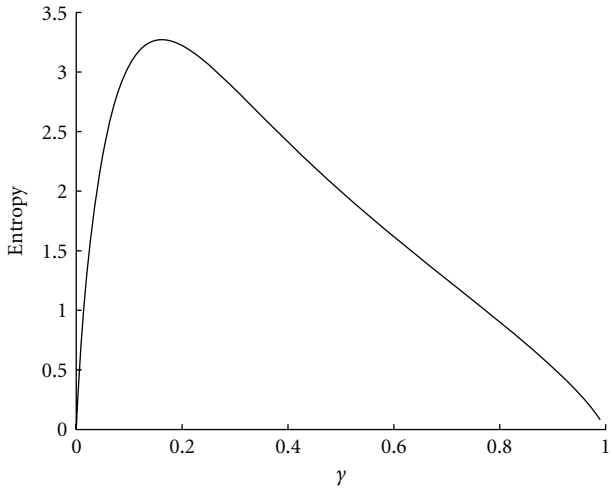


Figure 5: Entropy of the attacker as a function of sending announcement in the first round (γ). Number of nodes in one cluster: 10.

the number of cluster peers has been already balanced, this entropy can be calculated as follows:

$$\begin{aligned} H &= -\sum_{i=1}^{n+1} P_i \log P_i \\ &= -\sum_{i=1}^{n} \left((1-\gamma)^{i-1} \gamma \log \left((1-\gamma)^{i-1} \gamma \right) \right) \\ &\quad - (1-\gamma)^n \log (1-\gamma)^n, \end{aligned} \tag{5}$$

where γ is the probability of sending an announcement in the first round and n is the balanced number of cluster peers.

In Figure 5, we plotted formula (5). If γ is large, then the uncertainty of the attacker is low, because one of the first few senders will become the cluster aggregator with very high probability. If γ is very small then the uncertainty of the attacker is small again, because no cluster aggregator will be elected in the first round with high probability, and therefore, the first sender of the second round will be the cluster aggregator. The ideal γ value corresponds to the maximum entropy, which can be easily computed by the nodes locally from formula (5). For instance, Table 2 shows some ideal γ values for different number of nodes in one cluster. The fifth row ($H_{\max}$) shows the maximal entropy (uncertainty) that any kind of election protocol can achieve with the given number of nodes. This is achieved if every node is equiprobably elected from the viewpoint of the attacker. This value is closely approached by $H(\hat{\gamma})$, where $\hat{\gamma}$ is very close to the optimal solution (the difference between the found value and the optimal value can be arbitrarily small, and depends on the number of iterations the estimation algorithm uses). Using the found $\hat{\gamma}$ value, the order of the messages has no meaning for the attacker.

3.2.3. Node Capture Attacks. If an attacker can compromise a node, it can reveal some sensitive information, even when the system uses the local key-based protocol. If the compromised node is a cluster aggregator, then all the previously stored messages can be revealed. The attacker can decide to demolish the node, modify the stored values, simply use the captured data, or modify the aggregation functions.

If the compromised node is not a cluster aggregator, then the attacker can reveal the cluster aggregator of that node, which can result in the same situation described in the previous paragraph.

Table 2: Optimal γ values ($\hat{\gamma}$) for different number of nodes in one cluster. Achieved entropy ($H(\hat{\gamma})$) and maximal entropy ($H_{\max} = \log_2 n$).

n	10	25	50	100
$\hat{\gamma}$	0.167	0.082	0.049	0.027
$n\hat{\gamma}$	1.67	2.05	2.45	2.7
$H(\hat{\gamma})$	3.281	4.410	5.312	6.218
$H_{\max}$	3.322	4.644	5.644	6.644

3.3. Data Forwarding and Querying. The problem of forwarding the measured data to the aggregators without revealing the identity of the aggregators is a well-known problem in the literature, called anonymous routing [12–14].

Anonymous routing lets us route packets in the network without revealing the destination of the packet. A short overview of anonymous routing can be found in Section 5.

With anonymous routing any node can send the measurements to the aggregators without revealing the identity of it. An operator can query the aggregator with the help of an ordinary node which uses anonymous routing towards the aggregator.

Anonymous routing introduces significant overhead in the traffic. However this can be partially mitigated by synchronizing the data transmissions. Instead of suggesting such an approach, in this paper we elaborate a more challenging situation where the identity of the aggregators is unknown to the cluster members as well in Section 4.3. The clear advantage is that even if a node is compromised, its aggregator cannot be identified.

4. Advanced Protocol

The advanced private data aggregation protocol is designed to withstand the compromise of some nodes without revealing the identities of the aggregator. The protocol consists of four main parts. The first part is the initialization, which provides the required communication channel. The second part is needed for the data aggregator election. This subprotocol must ensure that the cluster does not remain without a cluster aggregator. This must be done without revealing the identity of the elected aggregator. The third part is needed for the data aggregation. This subprotocol must be able to forward the measured data to the aggregator without knowing its identifier. The last part must support the queries, where an operator queries some stored aggregated data.

Table 3: Summary of complexity of the advanced protocol. N is the number of nodes in the cluster.

	Election	Aggregation	Query
Message complexity	$O(N^2)$	$O(N)$	$O(N)$
Modular exponentiations	$4N$[1]	0	0
Hash computations	0	0	1

[1] 4 exponentiations for generating the two messages with knowledge proofs and 4N-4 exponentiations for checking the received knowledge proofs.

In the following, the description of each subprotocol follows the same pattern. First the goal and the requirements of the subprotocol are discussed, then the subprotocol itself is presented. After the presentation of the subprotocol, we analyze how it achieves its goal even in the presence of an attacker, and what data and services it provides for the next subprotocol.

At the end of this section, misbehavior is analyzed. We discuss, what an attacker can achieve, if its goal is not to identify the aggregators of the cluster, but to confuse the operation of the protocols.

In the following, it is assumed that every node knows which cluster it belongs to. The protocol descriptions are considering only one cluster, and separate instances of the protocol are run in different clusters independently.

The complexity of each subprotocol is summarized in Table 3.

4.1. Initialization. The initialization phase is responsible for providing the medium for authenticated broadcast communication. In the following, we shortly review the approaches of broadcast authentication in wireless sensor networks, and give some efficient methods for broadcast communication.

The initialization relies on some data stored on each node before deployment. Each node has some unique cryptographic credentials to enable authentication and is aware of the cluster identifier it belongs to. In the following, without further mentioning, we will assume, that each message contains the cluster identifier. Every message addressed to a cluster different from the one a node belongs to is discarded by the node. First, we briefly review the state of the art in broadcast authentication, then we propose a connected dominating set-based broadcast communication method, which fits well to the following aggregation and query phases.

4.1.1. Broadcast Authentication. Broadcast authentication enables a sender to broadcast some authenticated messages efficiently to a big number of potential receivers. In the literature, this problem is solved with either digital signatures or hash chains. In this section, we reviews some solutions from both approaches.

For the sake of completeness, Message Authentication Codes (MACs) must also be mentioned here [15]. MACs are based on symmetric cryptographic primitives, which enable very efficient computation. Unfortunately, the verifier of a MAC must also possess the same cryptographic credential the generator used for generating the MAC. It means that every node must know every credential in the network, to verify every message broadcast to the network. This full knowledge can be exploited by an attacker who compromises a node. The attacker can impersonate any other honest node, which means that if only one node is compromised, message authenticity can no longer be ensured.

One solution to the node compromise is the hop by hop authentication of the packets. In hop by hop authentication, every packets authentication information is regenerated by every forwarder. In this case, it is enough to only have a shared key with the direct neighbors of a node. In case of node compromise, only the node itself and the direct neighbors can be impersonated. Such a neighborhood authentication is provided by Zhu et al. in LEAP [9], where it is based on so-called cluster keys.

To make the authentication scheme robust against node compromise, one approach is the usage of asymmetric cryptography, namely, digital signatures.

Digital signatures are asymmetric cryptographic primitives, where only the owner of a private key can compute a digital signature over a message, but any other node can verify that signature. Computing a digital signature is a time-consuming task for a typical sensor node, but there exist some efficient elliptic curve-based approaches in the literature [16–19].

One of the first publicly available implementations was the TinyECC module written by Liu and Ning [16]. A more efficient implementation is the NanoECC module. Proposed by Szczechowiak et al. [17]. It is based on the MIRACL cryptographic library [20]. Up to now, to the best of our knowledge, the fastest implementations are the TinyPBC by Oliveira et al. [18], which is based on the RELIC toolkit [21], and the TinyPairing proposed by Xiong et al. in [19].

Another approach is proposed for broadcast authentication in wireless sensor networks by Perrig et al. in [22]. The μTESLA scheme is based on delayed release of hash chain values used in MAC computations. The scheme needs secure loose time synchronization between the nodes. The μTESLA scheme is efficient if it is used for authenticating many messages, but inefficient if the messages are sparse. Consequently, if only the rarely sent election messages must be authenticated, then the time synchronization itself can cause a heavier workload than simple digital signatures. If the aggregation messages must also be authenticated, then μTESLA can be an efficient solution. A DoS resistant version specially adapted for wireless sensor networks is proposed by Liu et al. in [23]. A faster but less secure modification is proposed by Huang et al. in [24].

In the following we will assume that an efficient broadcast authentication scheme is used without any indication.

4.1.2. Broadcast Communication. Broadcast communication is a method that enables sending information from one source to every other participant of the network. In wireless networks it can be implemented in many ways, like flooding the network or with a sequence of unicast messages.

A natural question would be, why broadcast communication is so important to the advanced protocol? The reason is that only broadcast communication can hide

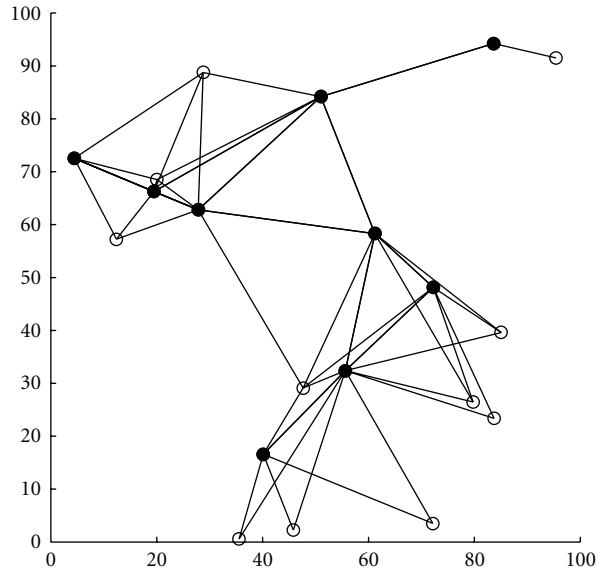


Figure 6: Connected dominating set. Solid dots represents the dominating set, and empty circles represent the remaining nodes. The connections between the non-CDS nodes of the network is not displayed on the figure.

the traffic patterns of the communication, thus not revealing any information about the aggregators.

An efficient way of implementing broadcast communication in wireless sensor networks is the usage of connected dominating set (CDS). The connected dominating set S of graph G is defined as a subset of G such that every vertex in $G - S$ is adjacent to at least one member of S, and S is connected. A graphical representation of a CDS can be found in Figure 6. The minimum connected dominating set (MDCS) is a connected dominating set with minimum cardinality. Finding an MDCS in a graph is an NP-Hard problem, however there are some efficient solutions which can find a close to minimal DCS in WSNs. For a thorough review of the state of the art of DCS in WSNs, the interested reader is referred to [25, 26].

In the following, we will assume that a connected dominating set is given in each cluster, and a minimum spanning tree is generated between the nodes in the CDS. Finding a minimum spanning tree in a connected graph is a well-known problem for decades. Efficient polynomial algorithms are suggested in [27, 28]. This kind of two layer communication architecture enables the efficient implementation of different kind of broadcast like communications, which are required for the following protocols. The spanning tree is used in the aggregation protocol in Section 4.3.

The simple all node broadcast communication can be implemented simply: if a node sends a packet to the broadcast address, then every node in the CDS forwards this message to the broadcast address. The CDS members are connected and every non-CDS member is connected to at least one CDS member by definition, so the message will be delivered to every recipient in the network. This approach is more efficient than simple flooding as only a subset of the nodes forwards the message, but the properties of the CDS ensures that every node in the cluster will eventually receive the broadcast information. Here, the notion of CDS parent (or simply parent) must be introduced. The CDS parent of node A is a node, which is in communication distance with A and is a member of the CDS.

The complexity of such a broadcast communication is $O(N)$, but actually it takes $|S|$ messages to broadcast some information, where $|S|$ is the number of nodes in the connected dominating set. If the CDS algorithm is accurate, than it can be very close to the minimum number of nodes required to broadcast communication.

In the following, we will use broadcast communication frequently to avoid that an attacker can gain some knowledge about the identity of the aggregators from the traffic patterns inside the network. Obviously we will not broadcast every message as it is in the network, because that would shortly lead to battery depletion and inoperability of the sensor network. Instead of automatically broadcasting every message, we will try to aggregate as much information as possible in each message to preserve energy. In the following sections, we will use the given CDS in different ways, and each particular usage will be described in the corresponding section.

The used communication patterns are closely related to and inspired by the Echo algorithm published by Chang in [29]. The Echo algorithm is a Wave algorithm [30], which enables the distributed computation of an idempotent operator in trees. It can be used in arbitrary connected graphs and generates a spanning tree as a side result.

4.2. Data Aggregator Election. The main goal of the aggregator node election protocol is to elect a node that can store the measurements of the whole cluster in a given epoch, but in such a way that the identity remains hidden. The election is successful if at least one node is elected. The protocol is unsuccessful if no node is elected, thus no node stores the data. In some cases, electing more than one node can be advantageous, because the redundant storage can withstand the failure of some nodes. In the following, we propose an election protocol, where the expected number of elected aggregators can be determined by the system operator, and the protocol ensures that at least one aggregator is always elected.

The election process relies on the initialization subprotocol discussed in Section 4.1. It requires an authenticated broadcast channel among the cluster members, which is exactly what the initialization part offers.

The election process consists of two main steps: (i) every node decides, whether it wants to be an aggregator, based on some random values. This step does not need any communication, the nodes compute the results locally. (ii) In the second step, an anonymous veto protocol is run, which reveals only the information that at least one node elected itself to be aggregator node. If no aggregator is elected, it will be clear for every participant, and every participant can run the election protocol again.

Step (i) can be implemented easily. Every node elects itself aggregator with a given probability p. The result of the election is kept secret, the participants only want to know that the number c of aggregators is not zero, without revealing the identity of the cluster aggregators. This is advantageous, because in case of node compromise, the attacker learns only whether the compromised node is an aggregator, but nothing about the identity or the number of the other aggregators. Let us denote the random variable representing the number of elected aggregators with C. It is easy to see that the distribution of C is binomial (N is the total number of nodes in one cluster):

$$\Pr(C = c) = \binom{N}{c} p^c (1-p)^{N-c}. \tag{6}$$

The expected number of aggregators after the first step is $c_E = Np$. So if on average $\hat{c}$ cluster aggregator is needed, then p should be $\hat{c}/N$ (this formula will be slightly modified after considering the results of the second step).

The probability that no cluster aggregator is elected is $(1-p)^N$.

To avoid the anarchical situation when no node is elected, the nodes must run step (ii) which proves that at least one node is elected as aggregator node, but the identity of the aggregator remains secret. This problem can be solved by an anonymous veto protocol. Such a protocol is suggested by Hao and Zieliński in [31].

Hao and Zieliński's approach has many advantageous properties compared to other solutions [32, 33], such as the property that it requires only 2 communication rounds.

The anonym veto protocol requires knowledge proofs. Informally, a knowledge proof allows a prover to convince a verifier that he knows a solution of a hard-to-solve problem without revealing any useful information about the knowledge. A detailed explanation of the problem can be found in [34].

A well known example of knowledge proof is given by Schnorr in [35]. The proposed method gives a noninteractive proof of knowledge of a logarithm without revealing the logarithm itself. The operation can be described briefly as follows. The proof of knowledge of the exponent of g_i^x consists of the pair $\{g^v, r = v - x_i h\}$, where $h = H(g, g^v, g_i^x, i)$ and H is a secure hash function. This proof of knowledge can be verified by anyone through checking whether g^v and $g^r g^{x_i h}$ are equal.

The operation of the anonym veto protocol consists of two consecutive rounds (G is a publicly agreed group with order q and generator g).

(1) First, every participant i selects a secret random value: $x_i \in \mathbb{Z}_q$. Then g_i^x is broadcast with a knowledge proof. The knowledge proof is needed to ensure that the participant knows x_i without revealing the value of x_i. Without knowledge proof, the node could choose g_i^x in a way to influence the result of the protocol (it is widely believed that for a given $g_i^x \pmod p$ it is hard to find $x_i \pmod p$, this problem is known as the discrete logarithm problem). Then every participant checks the knowledge proofs and computes a special product of the received values:

$$g^{y_i} = \frac{\prod_{j=1}^{i-1} g^{x_j}}{\prod_{j=i+1}^{N} g^{x_j}}. \tag{7}$$

(2) $g^{y_i c_i}$ is broadcast with a knowledge proof (the knowledge proof is needed to ensure that the node cannot influence the election maliciously afterwards). c_i is set to x_i for nonaggregators, while a random r_i value for aggregators.

The product $P = \prod_{i=1}^{N} g^{c_i y_i}$ equals 1 if and only if no cluster aggregator is elected (none vetoed the question: is the number of cluster aggregators elected zero?). If no aggregator is elected, then it will be clear for all participants, and the election can be done again. If P differs from 1, then some nodes are announced themselves to be cluster aggregators, and this is known by all the nodes.

If we consider the effect of the second step (new election is run if no aggregator is elected), the expected number of aggregators is slightly higher than in the case of binomial distributions. The expected number of aggregators are

$$c_E = \frac{Np}{1-(1-p)^N}. \tag{8}$$

The anonymity of the election subprotocol depends on the parts of the protocol. Obviously, the random number generation does not leak any information about the identity of the aggregator nodes, if the random number generator is secure. A cryptographically secure random number generator, called TinyRNG, is proposed in [36] for wireless sensor networks. Using a secure random number generator, it is unpredictable, who elects itself to be aggregator node.

The anonymity analysis of the anonym veto protocol can be found in [31]. The anonymity is based on the decisional Diffie-Hellman assumption, which is considered to be a hard problem.

The message complexity of the election is $O(N^2)$, which is acceptable as the election is run infrequently (N is the number of nodes in the cluster).

If this overhead with the 4 modular exponentiations (see Table 3 for the complexities and Table 1 for the estimated running times, note that RSA is based on modular exponentiation) is too big for the application, then it can use the basic protocol described in Section 3.1, where only symmetric key encryption is used.

In wireless sensor networks, the links in general are not reliable, packet losses occur in time to time. Reliability can be introduced by the link layer or by the application. As it is crucial to run the election protocol without any packet loss, it is required to use a reliable link layer protocol for this subprotocol. Such protocols are suggested in [37, 38] for wireless sensor networks.

As a summary, after the election subprotocol every node is equiprobably aggregator node. The election subprotocol ensures that at least one aggregator is elected and this node(s) is aware of its status. An outsider attacker does not know

the identity of the aggregators or even the actual number of the elected aggregator nodes. An attacker, who compromised one or more nodes, can decide whether the compromised nodes are aggregators, but cannot be certain about the other nodes.

4.3. Data Aggregation. The main goal of the WSN is to measure some data from the environment and store the data for later use. This section describes how the data is forwarded to the aggregator(s) without the explicit knowledge of the identifier(s) of the aggregator(s).

The data aggregation and storage procedure use the broadcast channel. If the covered area is so small or the radio range is so large that every node can reach each other directly, then the aggregation can be implemented simply. Every node broadcasts their measurement to the common channel, and the cluster aggregator(s) can aggregate and store the measurements. If the covered area is bigger (which is the more realistic case), a connected dominating set-based solution is proposed.

In each time slot, each ordinary node (not member of the CDS) sends its measurement to one neighboring CDS member (to the parent) by unicast communication. When the epoch is elapsed and all the measurements from the nodes are received, the CDS nodes aggregate the measurements and use a modification of the Echo algorithm on the given spanning tree to compute the gross aggregated measurement in the following way: each CDS member waits until all but one CDS neighbor sends its subaggregate to it, and after some random delay it sends the aggregate to the remaining neighbor. This means that the leaf nodes of the tree start the communication, and then the communication wave is propagated towards the root of the spanning tree. This behavior is the same as the second phase of the Echo algorithm. When one node receives the subaggregates from all of its neighbors, thus cannot send it to anyone, it can compute the gross aggregated value of the network. Then, this value is distributed between the cluster members by broadcasting it every CDS member.

This second phase is needed, so that every member of the cluster can be aware of the gross aggregated value, and the anonymous aggregators can store it, while the others can simply discard it. The stored data includes the time slot in which the aggregate was computed, and the environmental variables if more than one variable (e.g., temperature and humidity) is recorded besides the value itself.

The aggregation function can be any statistical function of the measured data. Some easily implementable and widely used functions are the minimum, maximum, sum, or average. In Figure 7, the aggregation protocol is visualized with five nodes and two aggregators using the average as an aggregation function.

The anonymity analysis of the aggregation subprotocol is quite simple. After the aggregation, every node possesses the same information as an external attacker can get. This information is the aggregated data itself, without knowing anything about the identity of the aggregators. If the operator wants to hide the aggregated data, it can use some techniques discussed in Section 5.

The message complexity of the aggregation is $O(N)$, where N is the number of nodes in the cluster. This is the best complexity achievable, because to store all the measurements by a single aggregator, all nodes must send the measurements towards the aggregator, which leads to $O(N)$ message complexity. In terms of latency, the advanced protocol doubles the time the aggregated measurement arrives to the aggregator compared to a naive system, where the identity of the aggregators is known to every participant. This latency is acceptable as in most WSN applications the time between the measurements is much longer than the time required to aggregate the data.

As mentioned in the election subprotocol, the protocol must be prepared to packet losses due to the nature of wireless sensor networks. In the aggregation subprotocol two kinds of packet loss can be envisioned: a packet can be lost before or after the final aggregate is computed. Both cases can be detected by timers and a resend request can be sent. If the resend is unsuccessful for some times, the aggregation must be run without those messages. If the lost message contains a measurement or subaggregate, then the final aggregate will be computed without that data leading to an inaccurate measurement. If the lost message contained the gross aggregate, then some nodes will not receive the gross aggregate. Here it is very useful that the network can have multiple aggregators, because if at least one aggregator received the data, the data can be queried by the operator.

4.4. Query. The ultimate goal of the sensor network is to make the measured data available to the operator upon request. While the aggregation subprotocol ensures that the measured data is stored by the aggregators, the goal of the query subprotocol is to provide the requested data to the operator and keep the aggregators' identity hidden at the same time.

One solution would be that the operator visits all the nodes and connects to them by wire. While this solution would leak no information about the identities of the aggregators to any eavesdropping attacker, the execution would be very time consuming and cumbersome. Moreover, the accessibility of some nodes may be difficult or dangerous (e.g., in a military scenario). Therefore, we propose a solution where it is sufficient for the operator to get in wireless communication range of any of the nodes. This node does not need to be an aggregator, as actually no one, not even the operator knows who the aggregator nodes are.

As a first step, the operator authenticates itself to the selected node O using the key k_O. After that, node O starts the query protocol by sending out a query, obtains the response to the query from the cluster, and makes the response available to the operator. In the following, we will assume that O is not a CDS node. (If it is indeed a CDS node, then the first and last transmission of the query protocol can be omitted.)

Node O broadcasts the query data Q with the help of the CDS nodes in the cluster. This is done by sending Q to the CDS parent, and then every CDS member rebroadcasts

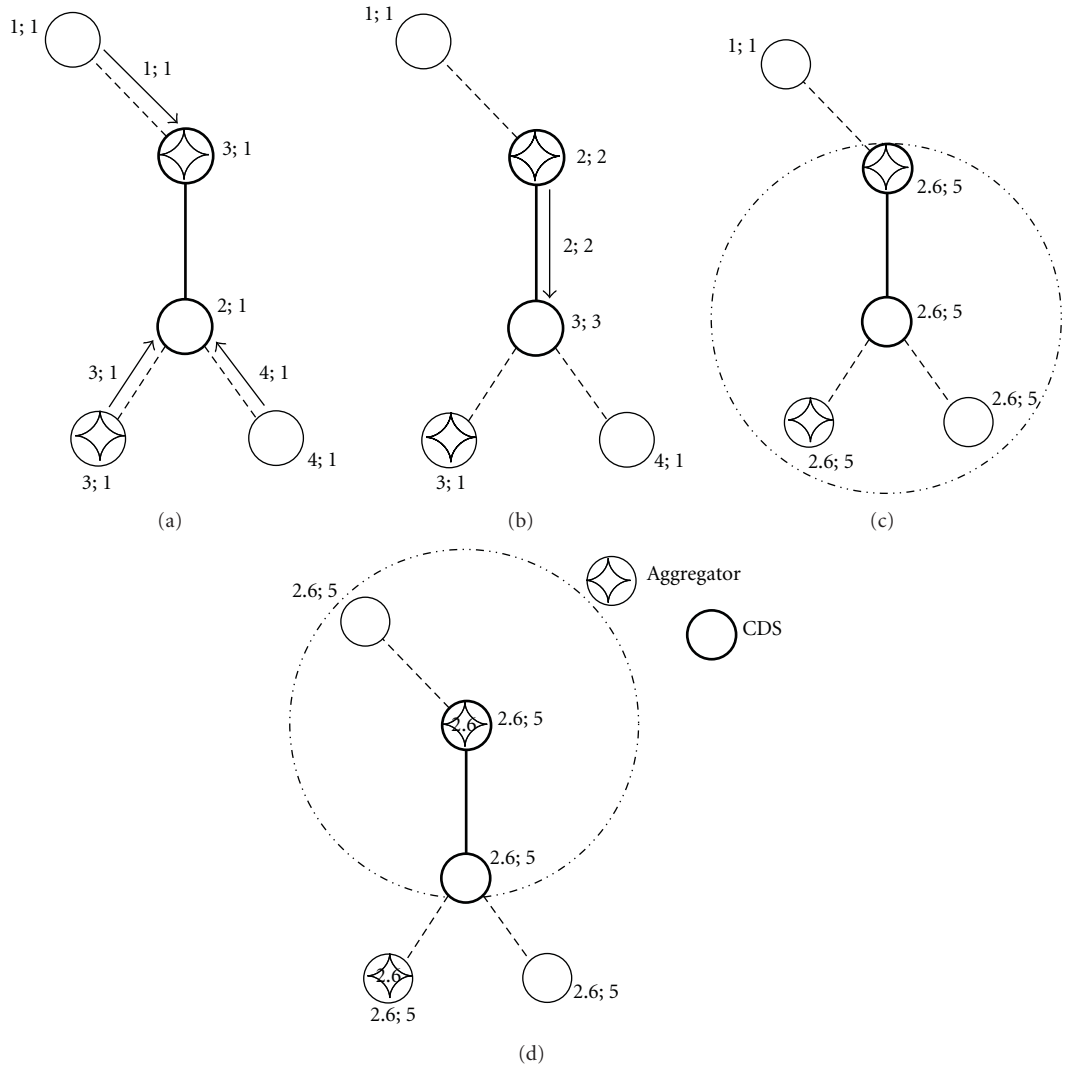


Figure 7: Aggregation example. The subfigures from left to right represents the consecutive steps of an average computation: (i) the measured data is ready to send. It is stored in a format of actual average; number of data. Non-CDS nodes sends the average to their parents. (ii) The CDS nodes start to send the aggregated value to its parents. (iii) A CDS node receives an aggregate from all of its neighbors, and starts to broadcast the final aggregated value. Nodes willing to store the value can do so. (iv) Other CDS nodes receiving the final value rebroadcasts it. Nodes willing to store the value can do so.

Q as it is received. The query Q describes what information the operator is interested in. It includes a variable name, a time interval, and a field for collecting the response to the query. It also includes a bit, called "aggregated", which will later be used in the detection of misbehaving nodes. For the details of misbehaving node detection, the reader is referred to Section 4.5; here we assume that the "aggregated" bit is always set meaning that aggregation is enabled.

The idea of the query protocol is that each node i in the cluster contributes to the response by a number R_i, which is computed as follows:

$$R_i = \begin{cases} h(Q \mid k_i), & \text{for nonaggregators,} \\ h(Q \mid k_i) + M, & \text{for aggregators,} \end{cases} \tag{9}$$

where M is the stored measurement (available only if the node is an aggregator), h is a cryptographic hash function, and k_i is the key shared by node i and the operator. Thus, nonaggregators contribute with a pseudorandom number $h(Q \mid k_i)$ computed from the query and the key k_i, which can later be also computed by the operator, while aggregator nodes contribute with the sum of a pseudorandom number and the requested measurement data. The sum is normal fix point addition, which can overflow if the hash is a large value.

The goal is that the querying node O receives back the sum of all these R_i values. For this reason, when the query Q is received by a non-CDS node from its CDS parent, it computes its R_i value and sends it back to the CDS parent in the response field of the query token. When a CDS parent receives back the query tokens with the updated response field from its children, it computes the sum of the received

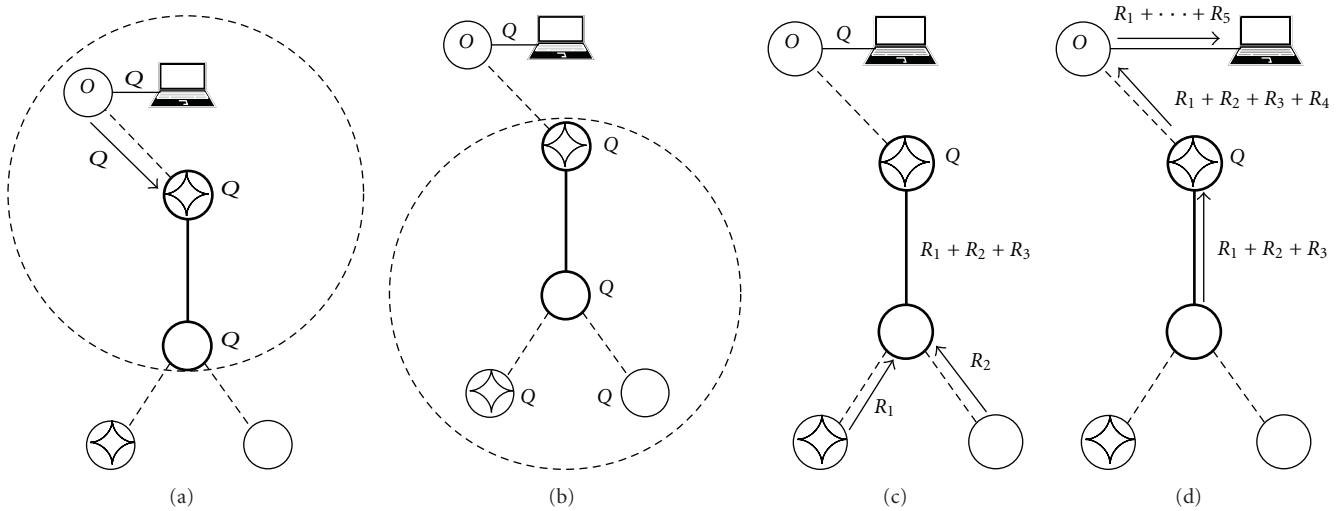


Figure 8: Query example. The subfigures from left to right represents the consecutive steps of a query: (i) the operator sends the Q query to node O. This node forwards it to its CDS parent. The CDS parent broadcasts the query. (ii) The CDS nodes broadcasts the query, so every node in the network is aware of Q. (iii) Every non-CDS node (except O) sends it response to its parent. (iv) The sum of the responses is propagated back to the parent of O (including the list of responding nodes, not on the figure), who forwards it to the operator through O.

R_i values and its own, and after inserting the identifiers of the nodes sends the result back to its parent. This is repeated until the query token reaches back to the CDS parent of node O, which can forward the response $R = \sum R_i$ and the list of responding nodes to node O, where the sum is computed by normal fix point addition. This operation is illustrated in Figure 8.

When receiving R from O, the operator can calculate the stored data as follows. First of all, the operator can regenerate each hash value $h(Q \mid k_i)$, because it stores (or can compute from a master key on-the-fly) each key k_i, and it knows the original query data Q. The operator can subtract the hash values from R (note that the responding nodes list is present in the response), and it gets a result $R' = cM$, where c is the actual number of aggregators in the cluster. (Note that each aggregator contributed the measurement M to the response, that is why at the end, the response will be c times M, where c is the number of aggregators.) Unfortunately, this number c is unknown to the operator, as it is unknown to everybody else. Nevertheless, if M is restricted to lie in an interval $[A, B]$ such that the intervals $[iA, iB]$ for $i = 1, 2, \ldots, N$ are nonoverlapping, then cM can fall only into interval $[cA, cB]$, and hence, c can be uniquely determined by the operator by checking which interval R' belongs to. Then, dividing R' with c gives the requested data M.

More specifically, and for practical reasons, the following three criteria need to be satisfied by the interval $[A, B]$ for our query scheme to work: (i) as we have seen before, for unique decoding of cM, the intervals $[iA, iB]$ for $i = 1, 2, \ldots, N$ must be nonoverlapping, (ii) in order to fit in the messages and to avoid integer overflow (In case of overflow, the result is not unique.), the highest possible value for cM, that is, NB must be representable with a prespecified number L, and (iii) it must be possible to map a prespecified number D of different values into $[A, B]$.

The first criterion (i) is met, if the lower end of each interval is larger than the higher end of the preceding interval:

$$0 < iA - (i-1)B = i(A-B) + B, \qquad i = 1, \ldots, N. \tag{10}$$

Note that if the above inequality holds for $i = N$, then it holds for every i, because $A - B$ is a negative constant and B is a positive constant. So it is enough to consider only the case of $i = N$:

$$\begin{aligned} 0 &< N(A-B) + B, \\ B &< \frac{N}{N-1} A. \end{aligned} \tag{11}$$

The second criterion (ii) means that

$$\begin{aligned} BN &< L, \\ B &< \frac{L}{N} \end{aligned} \tag{12}$$

while the third criterion (iii) can be formalized as

$$\begin{aligned} D &< B - A, \\ B &> A + D. \end{aligned} \tag{13}$$

Figure 9 shows an example for a graphical representation of the three criteria, where the crossed area represents the admissible (A, B) pairs. It can also be easily seen in this figure that a solution exists only if the B coordinate of the intersection of inequalities (11) and (13) meets criterion (12), or in other words

$$NM < \frac{L}{N}. \tag{14}$$

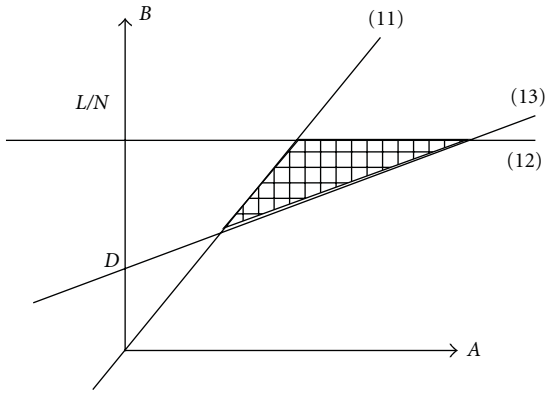


Figure 9: Graphical representation of the suitable intervals.

As a numerical example, let us assume that we want to measure at least 100 different values ($D = 99$); the microcontroller is a 16 bit controller ($L = 2^{16}$), and we have at most 20 nodes in each cluster ($N = 20$). Then a suitable interval that satisfies all three criteria would be $[A, B] = [2000–2100]$. Checking that this interval indeed meets the requirements is left for the interested reader. Finally, note that any real measurement interval can be easily mapped to this interval $[A, B]$ by simple scaling and shifting operations, and our solution requires that such a mapping is performed on the real values before the execution of the query protocol.

Our proposed protocol has many advantageous properties. First, the network can respond to a query if at least one aggregator can successfully participate in the subprotocol. Second, the operator does not need to know the identity of the aggregators, thus even the operator cannot leak that information accidentally (although, after receiving the response, the operator learns the actual number of the aggregator nodes). Third, the protocol does not leak any information about the identity of the aggregators: an attacker can eavesdrop the query information Q, and the R_i pseudo random numbers, but cannot deduce from them the identity of the aggregators. Finally, the message complexity of the query is $O(N)$, where N is the number of nodes in the cluster. This is the best complexity achievable, when the originator of the query does not know the identity of the aggregator(s). The latency of the query protocol depends on the longest path of the network rooted at node O.

As mentioned in the previous subprotocols, the protocol must be prepared to packet losses due to the nature of wireless sensor networks. Due to the packet losses, the final sum R is the sum of the responding nodes which is a subset of all nodes. That is why the identifiers must be included in the responses. The operator can calculate cM independently from the actual subset of responders. If at least one response from an aggregator gets to the operator, it can calculate M in the previously described way. If $cM = 0$, then it is clear for the operator that every aggregators' response is lost.

4.5. Misbehaving Nodes. In this section, we look beyond our initial goal. We briefly analyze what happens if a compromised node deviates from the protocol to achieve some goals other than just learning the identity of the aggregators.

In the election process, a compromised node may elect itself to be aggregator in every election. This can be a problem if this node is the only elected aggregator, because a compromised node may not store the aggregated values. Unfortunately this situation cannot be avoided in any election protocol, because an aggregator can be compromised after the election, and the attacker can erase the memory of that node. Actually our protocol is partially resistant to this attack, because more than one aggregator may be elected with some probability, and the attacker cannot be sure if the compromised node is the single aggregator node in the cluster.

During the aggregation, a misbehaving node can modify its readings, or modify the values it aggregates. The modification of others' values can be prevented by some broadcast authentication schemes discussed in Section 4.1.1. The problem of reporting false values can be handled by statistical approaches discussed in [39–41].

The most interesting subprotocol from the perspective of misbehaving nodes is the query protocol. In this protocol, a compromised node can easily modify the result of the query in the following way. A compromised node can add an arbitrary number X to the hash in (9) instead of using 0 or M. It is easy to see, that if X is selected from the interval $[A, B]$, then after subtracting the hashes, the resulting sum R' will be an integer in the interval $[(c + 1)A, (c + 1)B]$ (c is the actual number of aggregators, $c + 1$ nodes act like aggregators, the c aggregator, and the compromised node). A compromised node can further increase its influence by choosing X from the interval $[iA, iB]$. This means that the resulting sum R' will be in the interval $[(c + i)A, (ci)B]$. If X is not selected from interval $[jA, jB]$, $j = 1, \ldots, N$, then the result can be outside of the decodable intervals. This can be immediately detected by the operator (see Figure 10).

If the result is in a legitimate interval ($\exists j, R' \in [jA, jB]$), then the operator can further check the consistency by calculating $R' \bmod j$. If the result is zero, then it is possible, that no misbehaving node is present in the network. If the result is nonzero, the operator can be sure, that apart from the zeros and Ms, some node sent a different value, thus a misbehaving node is present in the network. It is hard for the attacker to guess j, because it neither knows the actual number of aggregators, nor can calculate R' from R by subtracting the unknown hashes.

If the modulus is zero, but the operator is still suspicious about the result, it can further test the cluster for misbehaving nodes with the help of the aggregated bit in the queries. This further testing can be done regularly, randomly, or on receiving suspicious results. If the aggregated bit is cleared in a query Q, then the CDS nodes doesnot sum the incoming replies, but forward them towards the agent O node as they are received. So if the operator wants to check if a misbehaving node is present in the network it can run a query Q with aggregated bit set, and then run the same query with cleared aggregated bit. If the two results are different, then the operator can be sure that a node wants to hide its malicious activity from the operator. If the two

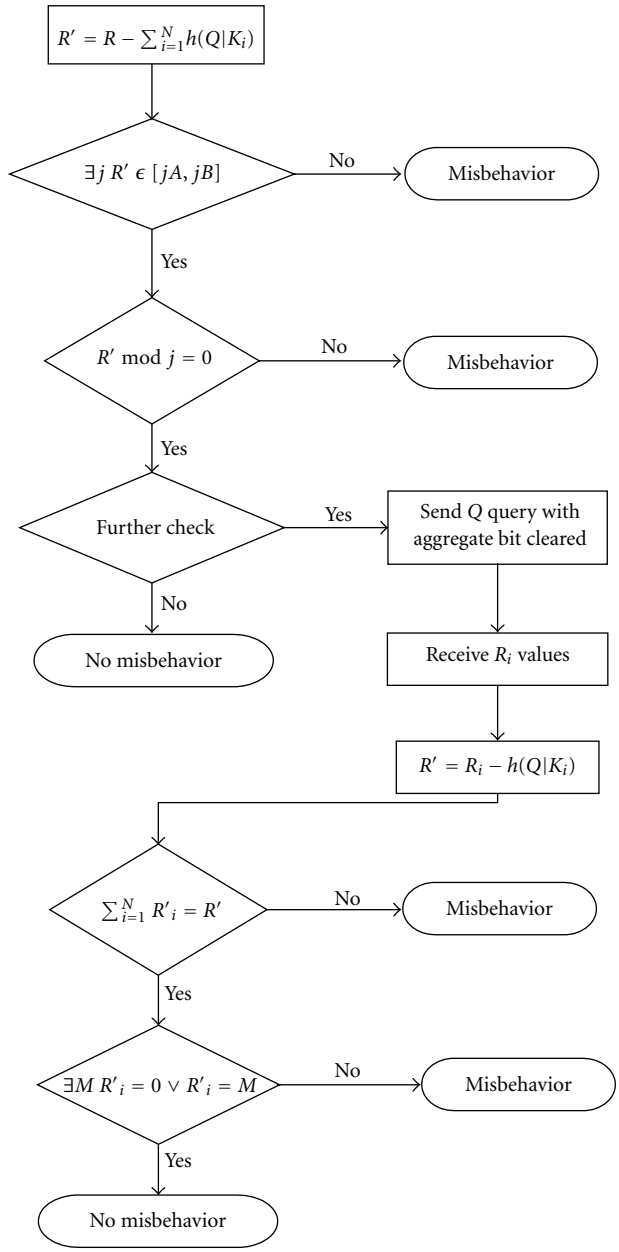


Figure 10: Misbehavior detection algorithm for the query protocol.

sums are equal, then the operator can further check the results from the second round. If the values are all equal after subtracting the hashes (not considering the zero values), then no misbehavior is detected, otherwise some node(s) misbehave in the cluster.

Note here that this algorithm does not find every misbehavior, but the misbehaviors not detected by this algorithm does not influence the operator. For example, two nodes can misbehave such that the first adds S to its hash and the second adds $-S$. It is clear that this misbehavior does not affect the result computed by the operator, because $S - S = 0$. Other misbehavior is not detected by the algorithm if a compromised non aggregator node sends M instead of 0. This is not detected by the algorithm, but doesnot modify the result the operator computes. The operation of misbehavior detection algorithm is depicted on Figure 10. This algorithm only detects if some misbehavior occurrs in the cluster, but does not necessarily find the misbehaving node. We left the elaboration of this problem for future work.

5. Related Work

A survey on privacy protection techniques for WSNs is provided in [42], where they are classified into two main groups: data-oriented and context-oriented protection. In this section, we briefly review these techniques, with an emphasis on those solutions that are closely related to our work.

In data-oriented protection, the confidentiality of the measured data must be preserved. It is also a research direction how the operator can verify if the received data is correct. The main focus is on the confidentiality in [43], while the verification of the received data is also ensured in [44].

According to [42] context-oriented protection covers the location privacy of the source and the base station. The source location privacy is mainly a problem in event-driven networks, where the existence and location of the event is the information, which must be hidden. The location privacy of the base station is discussed in [45]. The main difference between hiding the base station and the in-network aggregators is that a WSN regularly contains only one base station which is a predefined node, while at the same time there are more in network aggregators used in one network, and the nodes used as aggregators are periodically changed.

The problem of private cluster aggregator election in wireless sensor networks is strongly related to anonym routing in WSNs. The main difference between anonym routing and anonymous aggregation is that anonym routing supports any traffic pattern and generally handles external attackers, while anonymous aggregation supports aggregation-specific traffic patterns and can handle compromised nodes as well. In [12] an efficient anonymous on demand routing scheme called ARM is proposed for mobile ad hoc networks. For the same problem another solution is given in [13] (MASK), where a detailed simulation is also presented for the proposed protocol. A more efficient solution is given in [14], which uses low cryptographic overhead, and addresses some drawbacks of the two papers above. In [46] a privacy preserving communication system (PPCS) is proposed. PPCS provides a comprehensive solution to anonymize communication endpoints, keep the location and identifier of a node unlinkable, and mask the existence of communication flows.

The basis of this paper are two conference papers [1, 2]. Reference [1] describes the basic protocol discussed in Section 3, while [2] describes the advanced protocol discussed in Section 3. The main difference between this paper and [2], is that here a connected dominating set-based solution is proposed, while the previous paper assumed the existence of a Hamilton cycle inside the cluster. The management of such a cycle can be problematic in WSNs, while the CDS-based broadcast communication can be efficiently implemented in such a network [25]. Another contribution of this paper is the misbehavior detection algorithm, which solves a problem not discussed previously.

6. Conclusion

In wireless sensor networks, in-network data aggregation is often used to ensure scalability and energy efficient operation. However, as we saw, this also introduces some security issues: the designated aggregator nodes that collect and store aggregated sensor readings and communicate with the base station are attractive targets of physical node destruction and jamming attacks. In order to mitigate this problem, in this paper, we proposed two private aggregator node election protocols for wireless sensor networks that hide the elected aggregator nodes from the attacker, who, therefore, cannot locate and disable them. Our basic protocol provides fewer guarantees than our advanced protocol, but it may be sufficient in cases where the risk of physically compromising nodes is low. Our advanced protocol hides the identity of the elected aggregator nodes even from insider attackers, thus it handles node compromise attacks too.

We also proposed a private data aggregation protocol and a corresponding private query protocol for the advanced version, which allow the aggregator nodes to collect sensor readings and respond to queries of the operator, respectively, without revealing any useful information about their identity. Our aggregation and query protocols are resistant to both external eavesdroppers and compromised nodes participating in the protocol. The communication in the advanced protocol is based on the concept of connected dominating set, which suits well to wireless sensor networks.

In this paper we went beyond the goal of only hiding the identity of the aggregator nodes. We also analyzed what happens if a malicious node wants to exploit the anonymity offered by the system and tries to mislead the operator by injecting false reports. We proposed an algorithm that can detect if any of the nodes misbehaves in the query phase. We only detect the fact of misbehavior and leave the identification of the misbehaving node itself for future work.

In general, our protocols increase the dependability of sensor networks, and therefore, they can be applied in mission critical sensor network applications, including high-confidence cyber-physical systems where sensors and actuators monitor and control the operation of some critical physical infrastructure.

Disclosure

The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Acknowledgments

The work described in this paper is based on results of the WSAN4CIP Project (http://www.wsan4cip.eu), which receives research funding from the European Community's 7th Framework Programme. Apart from this, the European Commission has no responsibility for the content of this paper. The second author has also been supported by the Hungarian Academy of Sciences through the Bolyai János Research Fellowship. The authors are thankful to the anonymous reviewers for their useful comments and suggestions that helped to improve the quality of this paper.

References

[1] L. Buttyán and T. Holczer, "Private cluster head election in wireless sensor networks," in *Proceedings of the 6th IEEE*

International Conference on Mobile Adhoc and Sensor Systems (MASS '09), pp. 1048–1053, 2009.

[2] L. Buttyán and T. Holczer, "Perfectly anonymous data aggregation in wireless sensor networks," in *Proceedings of the 7th IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS '10)*, pp. 513–518, 2010.

[3] Y. R. Faizulkhakov, "Time synchronization methods for wireless sensor networks: a survey," *Programming and Computer Software*, vol. 33, no. 4, pp. 214–226, 2007.

[4] L. Buttyán and P. Schaffer, "Position-based aggregator node election in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2010, Article ID 679205, 15 pages, 2010.

[5] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings the 33rd Annual Hawaii International Conference on System Siences (HICSS '00)*, vol. 2, pp. 3005–3014, January 2000.

[6] R. Anderson and M. Kuhn, "Tamper resistance: a cautionary note," in *Proceedings of the 2nd USENIX Workshop on Electronic Commerce*, vol. 2, p. 1, USENIX Association, 1996.

[7] J. Lopez and J. Zhou, *Wireless Sensor Network Security*, Cryptology and Information Security Series, IOS Press, 2008.

[8] H. Chan, A. Perrig, and D. Song, "Random key predistribution schemes for sensor networks," in *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 197–213, IEEE Computer Society, May 2003.

[9] S. Zhu, S. Setia, and S. Jajodia, "LEAP: efficient security mechanisms for large-scale distributed sensor networks," in *Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS '03)*, pp. 62–72, ACM Press, October 2003.

[10] P. Ganesan, R. Venugopalan, P. Peddabachagari, A. Dean, F. Mueller, and M. Sichitiu, "Analyzing and modeling encryption overhead for sensor network nodes," in *Proceedings of the 2nd ACM International Workshop on Wireless Sensor Networks and Applications (WSNA '03)*, pp. 151–159, September 2003.

[11] K. Piotrowski, P. Langendoerfer, and S. Peter, "How public key cryptography influences wireless sensor node lifetime," in *Proceedings of the 4th ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN '06)*, pp. 169–176, October 2006.

[12] S. Seys and B. Preneel, "ARM: anonymous routing protocol for mobile ad hoc networks," in *Proceedings of the 20th International Conference on Advanced Information Networking and Applications (AINA '06)*, pp. 133–137, April 2006.

[13] Y. Zhang, W. Liu, W. Lou, and Y. Fang, "MASK: anonymous on-demand routing in mobile ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 9, pp. 2376–2385, 2006.

[14] T. Rajendran and K. V. Sreenaath, "Secure anonymous routing in ad hoc networks," in *Proceedings of the 1st ACM Bangalore Annual Conference (COMPUTE '08)*, ACM Press, New York, NY, USA, January 2008.

[15] B. Preneel and P. C. Van Oorschot, "On the security of iterated message authentication codes," *IEEE Transactions on Information Theory*, vol. 45, no. 1, pp. 188–199, 1999.

[16] A. Liu and P. Ning, "TinyECC: a configurable library for elliptic curve cryptography in wireless sensor networks," in *Proceedings of the 7th International Conference on Information Processing in Sensor Networks (IPSN '08)*, pp. 245–256, April 2008.

[17] P. Szczechowiak, L. B. Oliveira, M. Scott, M. Collier, and R. Dahab, "NanoECC: testing the limits of elliptic curve cryptography in sensor networks," in *Proceedings of the 5th European Conference on Wireless Sensor Networks (EWSN '08)*, vol. 4913 of *Lecture Notes in Computer Science*, pp. 305–320, 2008.

[18] L. B. Oliveira, M. Scott, J. López, and R. Dahab, "TinyPBC: pairings for authenticated identity-based non-interactive key distribution in sensor networks," in *Proceedings of the 5th International Conference on Networked Sensing Systems (INSS '08)*, pp. 173–180, IEEE, Kanazawa, Japan, June 2008.

[19] X. Xiong, D. S. Wong, and X. Deng, "TinyPairing: a fast and lightweight pairing-based cryptographic library for wireless sensor networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '10)*, April 2010.

[20] http://www.shamus.ie/.

[21] http://code.google.com/p/relic-toolkit/.

[22] A. Perrig, R. Canetti, J. D. Tygar, and D. Song, "The TESLA broadcast authentication protocol," *RSA CryptoBytes*, vol. 5, 2002.

[23] D. Liu, P. Ning, S. Zhu, and S. Jajodia, "Practical broadcast authentication in sensor networks," in *Proceedings of the 2nd Annual International Conference on Mobile and Ubiquitous Systems -Networking and Services (MobiQuitous '05)*, pp. 118–129, July 2005.

[24] Y. Huang, W. Ret, and K. Nahrstedt, "ChainFarm: a novel authentication protocol for high-rate any source probabilistic broadcast," in *Proceedings of the 6th IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS '09)*, pp. 264–273, October 2009.

[25] J. Blum, M. Ding, A. Thaeler, and X. Cheng, "Connected dominating set in sensor networksand manets," in *Handbook of Combinatorial Optimization*, D.-Z. Du and P. Pardalos, Eds., pp. 329–369, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[26] P. Jacquet, "Performance of connected dominating set in olsr protocol," Tech. Rep. RR-5098, INRIA, 2004.

[27] J. Kruskal and B. Joseph, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[28] R. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[29] E. J. H. Chang, "Echo algorithms: depth parallel operations on general graphs," *IEEE Transactions on Software Engineering*, vol. SE-8, no. 4, pp. 391–401, 1982.

[30] G. Tel, *Introduction to Distributed Algorithms*, Cambridge University Press, Cambridge, UK, 2nd edition, 2000.

[31] F. Hao and P. Zieliński, "A 2-round anonymous veto protocol," in *Proceedings of the 4th International Workshop on Security Protocols: Putting the Human Back in the Protocol*, vol. 5087 of *Lecture Notes in Computer Science*, pp. 202–211, Cambridge, UK, 2009.

[32] F. Brandt, "Efficient cryptographic protocol design based on distributed El gamal encryption," in *Proceedings of the 8th International Conference on Information Security and Cryptology (ICISC '05)*, vol. 3935 of *Lecture Notes in Computer Science*, pp. 32–47, 2006.

[33] D. Chaum, "The dining cryptographers problem: unconditional sender and recipient untraceability," *Journal of Cryptology*, vol. 1, no. 1, pp. 65–75, 1988.

[34] J. Camenisch and M. Stadler, "Proof systems for general statements about discrete logarithms," Tech. Rep., Department of Computer Science, ETH Zürich, Zürich, Switzerland, 1997.

[35] C. P. Schnorr, "Efficient signature generation by smart cards," *Journal of Cryptology*, vol. 4, no. 3, pp. 161–174, 1991.

[36] A. Francillon and C. Castelluecia, "TinyRNG: a cryptographic random number generator for wireless sensors network nodes," in *Proceedings of the5th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt '07)*, cyp, April 2007.

[37] A. Iqbal and S. A. Khayam, "An energy-efficient link layer protocol for reliable transmission over wireless networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 791201, 10 pages, 2009.

[38] C.-Y. Wan, A. T. Campbell, and L. Krishnamurthy, "PSFQ: a reliable transport protocol for wireless sensor networks," in *Proceedings of the ACM International Workshop on Wireless Sensor Networks and Applications*, pp. 1–11, ACM Press, 2002.

[39] L. Buttyán, P. Schaffer, and I. Vajda, "RANBAR: RANSAC-based resilient aggregation in sensor networks," in *Proceedings of the 4th ACM Workshop on Security of ad hoc and Sensor Networks (SASN '06)*, pp. 83–90, ACM Press, Alexandria, Va, USA, 2006.

[40] D. Wagner, "Resilient aggregation in sensor networks," in *Proceedings of the ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN '04)*, pp. 78–87, October 2004.

[41] L. Buttyán, P. Schaffer, and I. Vajda, "CORA: correlation-based resilient aggregation in sensor networks," *Ad Hoc Networks*, vol. 7, no. 6, pp. 1035–1050, 2009.

[42] N. Li, N. Zhang, S. K. Das, and B. Thuraisingham, "Privacy preservation in wireless sensor networks: a state-of-the-art survey," *Ad Hoc Networks*, vol. 7, no. 8, pp. 1501–1514, 2009.

[43] W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. Abdelzaher, "PDA: privacy-preserving data aggregation in wireless sensor networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 2045–2053, May 2007.

[44] B. Sheng and Q. Li, "Verifiable privacy-preserving range query in two-tiered sensor networks," in *Proceedings of the 27th IEEE Communications Society Conference on Computer Communications (INFOCOM '08)*, pp. 457–465, April 2008.

[45] J. Deng, R. Han, and S. Mishra, "Decorrelating wireless sensor network traffic to inhibit traffic analysis attacks," *Pervasive and Mobile Computing*, vol. 2, no. 2, pp. 159–186, 2006.

[46] H. Choi, P. McDaniel, and T. F. La Porta, "Privacy preserving communication in MANETs," in *Proceedings of the 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '07)*, pp. 233–242, June 2007.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 487209, 14 pages
doi:10.1155/2011/487209

Research Article

A Mobile-Beacon-Assisted Sensor Network Localization Based on RSS and Connectivity Observations

Guodong Teng,[1] Kougen Zheng,[2] and Ge Yu[1]

[1] *Institute of Service Engineering, Hangzhou Normal University, Hangzhou 310036, China*
[2] *College of Computer Science, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Kougen Zheng, zkg@zju.edu.cn

Received 15 February 2011; Revised 8 July 2011; Accepted 28 July 2011

Determining the physical positions of sensors has been a fundamental and crucial problem in wireless sensor networks (WSNs). Due to the inherent characteristics of WSNs, extremely limited resources available at each low-cost and tiny sensor node, connectivity-based range-free solutions could be a better choice to feature a low overall system cost. However, localization by means of mere connectivity may underutilize the proximity information available from neighborhood sensing. Although received signal strength (RSS) values are irregular and highly dynamic, it might provide heuristic information about which neighboring nodes are closer and which are further. We propose a distributed Mobile-beacon-assisted localization scheme based on RSS and Connectivity observations (MRC) with a specific trajectory in static WSNs. To ensure that MRC performs as true to reality as possible, we propose two improved approaches based on MRC to consider irregular radio scenario in the noisy environment. Comparing the performance with three typical range-free localization methods in static WSNs, our lightweight MRC algorithm with limited computation and storage overhead is more suitable for very low-computing power sensor nodes that can efficiently outperform better with only basic arithmetic operations in the simulation and physical environments.

1. Introduction

Wireless sensor networks (WSNs) have gained worldwide attention in recent years, particularly with the proliferation in micro-electro-mechanical systems (MEMSs) technology which has facilitated the development of smart sensors [1]. Determining the physical positions of sensors has been a fundamental and crucial problem in WSNs; for example, numerous location-dependent applications with sensor networks are proposed, such as habitat monitoring, battle-field surveillance, and enemy tracking. In addition, some of the routing protocols and network management mechanisms proposed for such networks are built on the assumption that geographic parameters of each sensor node are available [2].

Several schemes have been proposed for dealing with the sensor node localization. Based on whether accurate ranging is conducted at resource-constrained sensor node, the current approaches mainly fall into two categories: range-based and range-free. First, range-based approaches assume that sensor nodes are able to measure the distances, or even relative directions of their neighbor nodes, based on time of arrival (TOA), time difference of arrival (TDOA), angle of arrival (AOA), and received signal strength (RSS) technologies, and so forth. The range-based schemes typically have higher location accuracy but require additional hardware to measure distances or angles. Second, range-free approaches localize sensor nodes based on simple sensing, such as anchor proximity, connectivity-based, or localization events detection. Although the range-free schemes accomplish as high precision as the range-based ones, they provide an economic approach.

Due to the inherent characteristics of WSNs, extremely limited resources available at each low cost and tiny sensor node, connectivity-based solutions in range-free approaches could be a better choice to feature a low overall system cost. However, localization by means of mere connectivity may underutilize the proximity information available from neighborhood sensing [2]. A popular and widely used neighborhood sensing technique, always available in many sensor platforms such as Mica2, MicaZ, and TelosB, is RSS which is usually used in the range-based localization. Although RSS values are irregular and highly dynamic, the difference in RSS values consistently reflect the contrast of

physical distances [3]; that is, RSS might provide heuristic information about which neighboring nodes are closer and which are further [2].

The experiment results [2] have confirmed that network-wide monotonic relationship between RSS and physical distance does not hold. For example, at the receiver side, RSS sensing results are sensitive to small variance among different nodes [2]. However, the experiment results [2] have also confirmed that the monotonic RSS-distance relationship holds much well in the case of one sensor node. In addition, using a single mobile beacon that knows its position is broadly equivalent to using many static anchors each broadcasting once. These two facts motivate us to believe that a single mobile beacon which can travel the entire deployment region based on some traverse route can be used to help localize the entire network by connectivity observations and RSS relationship between the unknown node and the mobile beacon.

In this paper, we propose a distributed Mobile-beacon-assisted localization scheme based on RSS and Connectivity observations (MRC) with a specific trajectory in static WSNs. After the unknown nodes are deployed in a fixed-size area, a mobile beacon traverses the area while broadcasting packets. Then, the unknown nodes continuously observe the connectivity information from the mobile beacon. As a result, the unknown nodes' locations are bounded in the beacon's transmission area. Next, the unknown nodes utilize the comparison result of RSS from any two contacted neighbor beacons to narrow the location region of unknown node. To this end, the accuracy can be further improved when the unknown node obtains more comparison results from the connected beacon positions.

Furthermore, we also propose two improved approaches based on MRC, called MRC_Nearest and MRC_Centroid to consider the irregular radio scenario in the noisy environment. To ensure that MRC performs as true to reality as possible, four possible connectivity scenarios to the unknown node are further considered in the more general radio model. Evaluations show that the MRC_Centroid algorithm is robust and performs well even with severe interference.

This paper offers the following three major contributions:

(i) Improved Connectivity Restriction. we give a mobile beacon's trajectory for MRC in which some remarks regarding necessary properties are considered. Its novelty lies in combining 1-hop and 2-hop beacon position restrictions (connectivity restriction) at the same time;

(ii) Further Neighborhood Sensing Restriction. we utilize the proximity information available from neighborhood sensing as distance-related information (neighborhood sensing restriction) to further improve the precision of connectivity-based range-free localization in the mobile-beacon-assisted pattern;

(iii) Lightweight Computation and Storage. any unknown node only requires lightweight computation to solve the translation of position and limited storage overhead to retain (mainly 3 or 4) RSS and connectivity observations information whether or not the radio model is regular.

Comparing the performance with three typical range-free localization methods, the existing classic range-free localization algorithms with multiple static anchors, a single mobile-beacon-assisted range-free localization algorithm, and RSS-assisted range-free localization in static WSNs, our lightweight MRC algorithm with limited computation and storage overhead is more suitable for very low-computing power sensor nodes that can efficiently perform only basic arithmetic operations. Evaluation results show that MRC outperforms better in the simulation and physical environments.

The rest of this paper is organized as follows: Section 2 gives an overview of related works. Section 3 presents details of the proposed MRC, MRC_Nearest, and MRC_Centroid algorithms. Sections 4 and 5 show and discuss the simulation and physical evaluation results, respectively. Finally, Section 6 concludes our works.

2. Related Works

There has been a significant research activity in the area sensor network localization. Based on whether the range measurements are used at the resource-constrained sensor nodes, most of the existing approaches about sensor node localization fall into two categories: (1) range-based and (2) range-free localization. In addition, the sensor networks already incorporate GPS-equipped mobile beacon as part of the design, which can be a cost-effective way of achieving sensor node localization. In this section, we provide a brief literature review of these localization approaches.

Range-based approaches compute per-node location information iteratively or recursively based on measured distances among unknown nodes and a few anchors which precisely know their locations. Many range-based methods use techniques such as TOA [4, 5], TDOA (e.g., Cricket [6], AHLos [7], TPS [8]), AOA (e.g., APS [9], SpinLoc [10]), and RSS (e.g., Radar [11], wMDS [12], IndoorGPS [13], Sequence [14], Ranking [15]) to measure distance or angles among unknown nodes and anchors. Although range-based solutions can be suitably used in small-scale indoor environments, they are considered less cost effective for large-scale deployments.

Range-free approaches do not require accurate distance measurements but localize the unknown node based on anchor proximity (e.g., Centroid [16], APIT [17]), network connectivity information (e.g., DV-hop [18], MDS-MAP [19], SDP [20]), or localization events detection (e.g., Lighthouse [21], Spotlight [22]). Recently, some RSS-assisted range-free approaches are proposed, such as RSSL [23], PI [3], and RSD [2]. Instead of using absolute RSS to estimate the distance between two nodes with ordinary hardware, by contrasting the measured RSS values from the mobile beacon to a sensor node, RSSL and PI utilize the variance of RSS to estimate the unknown node position. Observing that per-node monotonic RSS-distance relationship holds well,

RSD captures a relative distance between 1-hop neighboring nodes from their neighborhood orderings to achieve better positioning accuracy than connectivity alone. For range-free connectivity-based localizations such as MDS-MAP and DV-hop, applying RSD is easy. Inspired by the empirical data results "Network-wide monotonic relationship between RSS and physical distance does not hold, but per-node monotonic RSS-Distance relationship holds well" shown in the RSD methods, but we propose a totally different range-free localization solution.

In addition, much research has been done on mobile-beacon-assisted localization for WSNs. The mobile beacon travels through the deployment area while broadcasting its location along the way. Unknown nodes localize themselves by monitoring range or connectivity information coming from the beacon. A general survey about mobile-beacon-assisted approaches can be found in [24]. Here, we review some works closely related to our proposed algorithms.

The paper [25] proposed some static paths to guide the GPS-equipped mobile beacon to achieve better coverage, better accuracy, and shorter path length, which is consistent with our design goals of trajectory. Significantly different from these static paths which are designed for range-based sensor network localization, such as RSS, TOA, and AOA under most common noise models (mostly Gaussian), our trajectory is designed for one of range-free approaches, based on RSS and connectivity observation.

In the range-free, distributed, and probabilistic algorithms MSL [26] and MSL*, each sensor node uses observations from only those neighbors (1-hop) that have better location estimates than it; that is, the weight of a sample is determined using the neighbors' observations besides the location announcements from the beacons. MSL modified the mobility model to allow this algorithm to work in static WSNs. MBL [24] relies on direct arriver and leaver information (2-hops) from a single mobile-assisted beacon which outperforms MSL when both of them use only a single mobile beacon with random trajectory for localization in static WSNs. All of these methods (MSL, MBL, and their improved versions) use the particle filter [27] (also called sequential Monte Carlo method) to perform a probabilistic localization on a sample representation. However, particle filters approach is quite power consuming in the tiny sensor nodes.

Our proposed methods differ significantly from those previous mobile-beacon-assisted range-free localization works because we utilize a mobile beacon with a specific trajectory and observe both the neighborhood sensing and connectivity observations (1-hop and 2-hop restrictions) with lightweight computation and storage overhead to improve precision in the localization process.

3. Design

3.1. MRC Localization Scheme. In this section, we describe our novel Mobile-beacon-assisted localization based on RSS and Connectivity observations, which we call MRC. Let us consider a sensor network with M static sensor nodes in a 2D plane which do not have a priori known locations (called unknown nodes) and a single mobile node (called beacon or anchor), equipped with localization hardware, for example, GPS, which allows it to know its location at all times. Here, we suppose that each sensor (unknown node or beacon) has the same ideal radio range r. In the regular radio model, the transmission range is equal to radio range, and, in the irregular radio model, the transmission range will be influenced by the environment noise. Any two sensors are called neighbors (or neighborhood) if the distance between them is less than or equal to the one-hop transmission range, otherwise known as unneighbor. After random deployment of the unknown nodes in a fixed-size area, the beacon traverses the entire sensor network with a specific trajectory while broadcasting packets which contain the coordinates of its own and some other related information. At any time, every unknown node within the radio range of the mobile beacon will hear a location announcement (called connectivity observation) from that beacon. We do not assume very tightly synchronized clocks. In a realistic deployment, it would be necessary to deal with network collisions and account for missed messages [28]. Location estimates and observations are assumed to be available at discrete times. For dynamic state estimation, the discrete-time approach is widespread and convenient [27].

Any unknown node receiving the packets from the mobile beacon can recognize that it is in the area around the beacon's current location with a certain probability. With each connectivity observation in a series of different times, the unknown node's location is bounded in the beacon's transmission area. The accuracy can be improved when the unknown node obtains more connectivity observations from the beacon. In addition, the unknown node further obtains the geometrical relationship of a series of beacon positions by contrasting the measured power of the signal at the receiver. Namely, after adopting the connectivity observation, the unknown node utilizes the comparison result of RSS derived from any two neighbor beacon positions to narrow the estimated location region of unknown node. As a result, the precision of position can be improved when the unknown node obtains more comparison results from the neighbor beacons.

Figure 1 shows an example situation. We supposed that the point A, B, and C shown in Figure 1 is a series of positions of mobile beacon passed. The unknown node could derive nonneighbor beacon positions by its neighbor beacon positions. This assumption will hold as long as the unknown node obtains the trajectory of mobile beacon. Thus, the unknown node S knows the beacon position A, B, and C in Figure 1. The unknown node S obtains the connectivity observation from the beacon positions A and B and does not receive the packets from the beacon position C. Then, the unknown node knows it is within distance r of beacon positions A and B but must not be located within the distance r of beacon position C; that is, the position of the unknown node S is within the shaded region in Figure 1. Next, the unknown node S compares the RSS values obtained from beacon positions A and B. The comparison result shows that the RSS value obtained from position A is larger than B, $SA > SB$; that is, the beacon position A is closer than

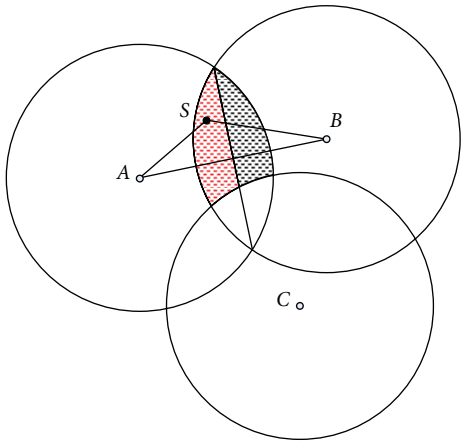


Figure 1: MRC localization scheme.

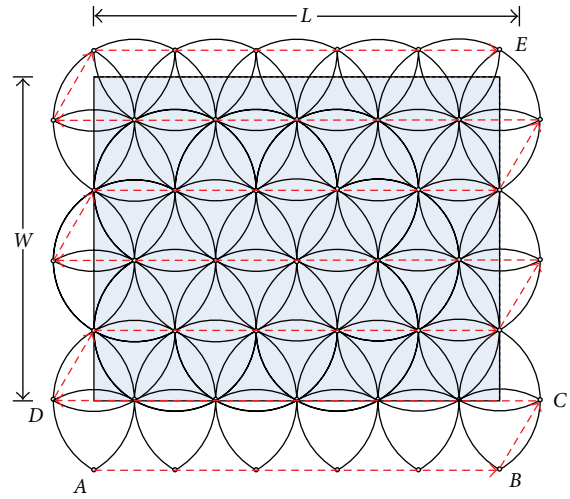


Figure 2: Trajectory of mobile beacon.

the beacon position B from the unknown node position S, $|AS| < |BS|$, where SA and SB denote the corresponding RSS values and $|AS|$ and $|BS|$ denote the corresponding distances. It is straightforward to show that the red-shaded region in Figure 1 enumerates all the possible estimated locations of the unknown node S.

If the measured two RSSs are the same or very close, that is, the difference between these two RSS values is less than a threshold, then the unknown node takes the center of these two beacon positions as its estimated position.

3.2. Trajectory of Mobile Beacon. The mobile-assisted localization scheme faces a new problem, "What is the optimum beacon route (as the same concept trajectory in our paper, beacon trajectory in [29], movement strategy in [30], movement pattern in [23], and movement path in [31])?" [29] or called trajectory selection problem. Notice that the problem is quite difficult. On one hand, the position of the unknown node is not known a priori [29]. Then, the trajectory of mobile beacon cannot be designed for the specific distribution of unknown nodes in the deployment area. On the other hand, the movement strategy of mobile beacon should be related to the specific localization scheme. Namely, the optimum trajectory in one mobile-beacon-assisted localization scheme is not necessarily optimal in the others. To this end, the trajectory of the mobile beacon to locate all the unknown nodes needs further considerations.

In this section, we will make some remarks regarding the following characteristics that the trajectory for MRC should have.

First, to improve the average position accuracy of unknown nodes, the trajectory of mobile beacon should cover the entire unknown nodes deployment area; that is, all the unknown nodes should be located when the mobile beacon completes all the paths.

Second, due to the limited resources available in the tiny sensor nodes, the trajectory of mobile beacon should be shorter, so as to achieve shorter location time and consume smaller energy cost. In addition, the unknown node should derive the unneighbor beacon position from such limited number of known neighbor beacon positions, easily, to further narrow the bounded transmission area obtained from neighbors' connectivity observation.

Third, the accuracy can be further improved when the unknown node contacts more neighbor beacons. However, the computational complexity of the unknown node also increases. Thus, the trajectory of mobile beacon should obtain a tradeoff between the localization accuracy and the computational complexity.

Therefore, let us consider the above characteristics to give the following trajectory of the mobile beacon shown in Figure 2, in which red-dotted line denotes the trajectory of mobile beacon. Figure 2 shows the length L and the width W of rectangular sensor nodes deployment area. The mobile beacon moves along the direction of the arrow in the red-dotted line, and the small circles denote a series of beacon positions. Namely, the mobile beacon starts from position A, through B, C two positions, to position D, then continues moving along the red-dotted line, and finally to position E. The distance between any two adjacent beacon positions of movement is r; that is, the minimum distance between the transmitted beacon points is equal to the radio range. This configuration of distance can obtain a tradeoff between the localization accuracy and the length of trajectory as shown in the latter evaluations. The arc in Figure 2 denotes the radio range of the corresponding beacon position.

This movement pattern for MRC meets the previously mentioned remarks. First, the rectangle area in Figure 2 represents the largest possible area of unknown node deployment. Any unknown node in this area can contact the connectivity information from the beacon 3 or 4 times; that is, all possible positions of unknown nodes are fully covered by at least 3 or 4 noncollinear beacons. Second, trajectory length is much smaller than the random trajectory in MBL and is very close to some classic trajectory, such as movement pattern in ADO [23], optimal trajectory in PI [3]. In addition, the unknown node could judge the unneighbor beacon' positions from these 3 or 4 contacted neighbor

beacons's position, and these unneighbor beacon positions could be used as 2-hop noncontacted position information. Third, the localization precision of MRC with this movement pattern is much higher as shown in latter evaluations. In addition, each unknown node only observes 3 or 4 times connectivity information and compares limited RSS values.

3.3. Regular Radio Model. The MRC algorithm can be broken down into three stages: (1) beacon acquisition, (2) region determination, and (3) location calculation. These stages are performed at individual unknown node in a purely distributed fashion.

(1) Beacon Acquisition. In the first stage, the mobile beacon is traversing the specified location in the unknown node deployment area as the trajectory described in Section 3.2. At each specified position, the mobile beacon broadcasts packets which contain the coordinates of its own and some other related information. When the unknown node itself is within the radio range of the mobile beacon's current specified position, the unknown node will obtain the connectivity observations from the beacon and measure the corresponding RSS values. The unknown node retains this information in the memory until the end of localization.

(2) Region Determination. When the mobile beacon with the trajectory finishes moving, any unknown node will contact the mobile beacon three, four, or six times in the rectangular deployment area. Without loss of generality, we show the first two cases in Figure 3.

Case 1 (3-connectivity). If the unknown node only contacts 3 times with the mobile beacon throughout mobile beacon moving process, the position of unknown node is within the overlapping regions of the transmission area of these three contacted mobile beacon positions and out of any other radio range of beacon positions. We call this case of relationship between the unknown node and beacon positions 3-connectivity.

As shown in Figure 3, the unknown node S is in the radio range of beacon positions A, B, and C and out of any other beacon positions, such as D, E, and F. Then, the unknown node is within the area formed by Region (1)~Region (6) shown in Figure 3. The advantage of 3-connectivity case to unknown node is the ability to combine 1-hop and 2-hop beacon position restrictions at the same time; that is, the unknown node with 3-connectivity case is within 1-hop neighbor beacon positions (A, B, and C) and out of 2-hop beacon positions (D, E, and F).

Once the 3-connectivity case happens, the RSS values from these three beacon positions A, B, and C in the unknown node have been measured, and the unknown node obtains the value SA, SB, and SC, respectively.

If $SA < SB$ and $SA < SC$ and $SB < SC$, then the unknown node S is within the Region (1).

If $SA > SB$ and $SA < SC$ and $SB < SC$, then the unknown node S is within the Region (2).

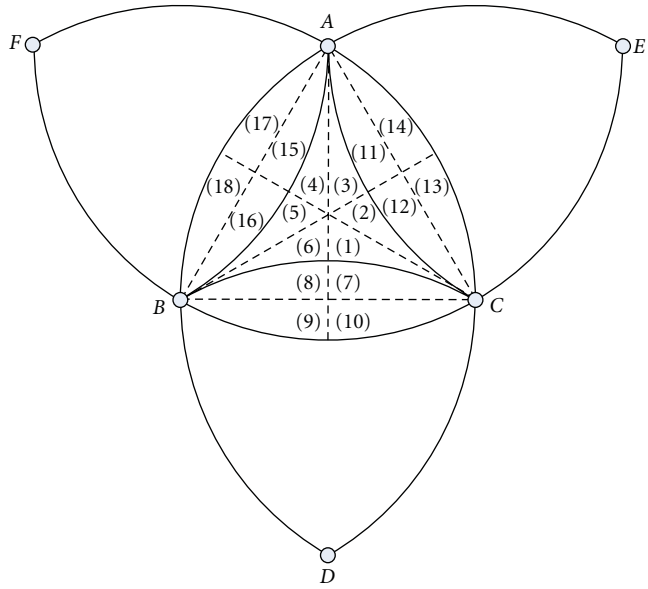


Figure 3: Different location regions.

If $SA > SB$ and $SA > SC$ and $SB < SC$, then the unknown node S is within the Region (3).

If $SA > SB$ and $SA > SC$ and $SB > SC$, then the unknown node S is within the Region (4).

If $SA < SB$ and $SA > SC$ and $SB > SC$, then the unknown node S is within the Region (5).

If $SA < SB$ and $SA < SC$ and $SB > SC$, then the unknown node S is within the Region (6).

Thus, after adopting the 3-connectivity observation with 1-hop and 2-hop restrictions, the unknown node utilizes the comparison result of RSS to further narrow the estimated location region of unknown node.

Case 2 (4-connectivity). If the unknown node just contacts 4 times with mobile beacon throughout mobile beacon moving process, the position of unknown node is within the overlapping regions of the transmission area of these four contacted mobile beacon positions and out of any other radio range of beacon positions. We call this case of relationship between the unknown node and beacon positions 4-connectivity.

As shown in Figure 3, the unknown node S is within one of 4 beacon positions (e.g., $ABCD$, $ABCE$, or $ABCF$) transmission area, and out of any other beacon positions. Without loss of generality, let the unknown node be within the radio range of beacon positions A, B, C, and D. Then, the unknown node is within the area formed by Region (7)~Region (10).

Once the 4-connectivity case happens, the RSS values from four beacon positions A, B, C, and D in the unknown node have been measured, and the unknown node obtains the value SA, SB, SC, and SD, respectively.

If $SA > SD$ and $SB < SC$, then the unknown node S is within the region (7).

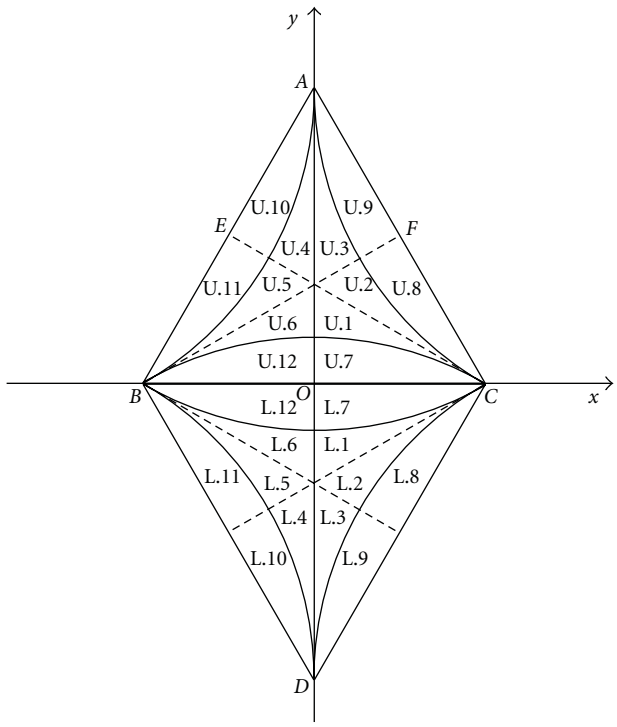


Figure 4: Relative regions.

If $SA > SD$ and $SB > SC$, then the unknown node S is within the region (8).

If $SA < SD$ and $SB > SC$, then the unknown node S is within the region (9).

If $SA < SD$ and $SB < SC$, then the unknown node S is within the region (10).

Case 3 (7-connectivity). If the unknown node S receives seven different beacon positions connectivity information, then S is on the trajectory of the mobile beacon. This special case is called 7-connectivity. It is obvious that S is uniquely determined by such seven beacon positions. S will take the centroid of seven connected beacon positions as its position.

(3) Location Calculation. After the position of unknown node is restricted to a particular region in the previous stage, the unknown node will take the centroid of restricted region as its estimated position.

Case 1 (3-connectivity). The triangle formed by three beacon positions in the 3-connectivity case is either Upper triangle or Lower triangle, where Upper triangle denotes the x-coordinate of one beacon position larger than the other two beacon positions' x-coordinate. As shown in Figure 4, the triangles ABC and BDC belong to Upper triangle and Lower triangle, respectively. We denote different restricted region by U.1~U.6 in the Upper triangle and L.1~L.6 in the Lower triangle. If we can obtain the centroid of region U.1 in Figure 4, any other centroid of region in the Upper and Lower triangles will be determined by the geometrical symmetry.

Now, we calculate the centroid of restricted region U.1. We denote the horizontal x and vertical y of region U.1 centroid $u1$ by $u1 = (u1 \cdot x,\ u1 \cdot y)$. As shown in Figure 4, we assume that the midpoint of line segment BC is located in the origin of coordinate O and $|BC| = r$. From the definition of centroid, the geometric center of the object's shape, we obtain

$$u1 \cdot x = \frac{\int_0^{r/2} \int_{\sqrt{r^2-x^2}-(\sqrt{3}/2)r}^{-(\sqrt{3}/3)x+(\sqrt{3}/6)r} x\,dy\,dx}{\int_0^{r/2} \int_{\sqrt{r^2-x^2}-(\sqrt{3}/2)r}^{-(\sqrt{3}/3)x+(\sqrt{3}/6)r} dy\,dx},$$
$$u1 \cdot x = \frac{\int_0^{r/2} \int_{\sqrt{r^2-x^2}-(\sqrt{3}/2)r}^{-(\sqrt{3}/3)x+(\sqrt{3}/6)r} y\,dy\,dx}{\int_0^{r/2} \int_{\sqrt{r^2-x^2}-(\sqrt{3}/2)r}^{-(\sqrt{3}/3)x+(\sqrt{3}/6)r} dy\,dx}. \quad (1)$$

Figure 4 shows that $u2$ and $u1$ are two centroid points which are symmetric with respect to a given line segment CE. The equation of the line through points C and E is given by

$$\left(2\sqrt{3}\right)x + 6y - \sqrt{3}r = 0. \quad (2)$$

Thus,

$$u2 \cdot x = u1 \cdot x - \frac{2 * (2\sqrt{3}) * ((2\sqrt{3}) * u1 \cdot x + 6 * u1 \cdot y - \sqrt{3}r)}{\sqrt{\left((2\sqrt{3})^2 + 6^2\right)}},$$
$$u2 \cdot y = u1 \cdot y - \frac{2 * 6 * ((2\sqrt{3}) * u1 \cdot x + 6 * u1 \cdot y - \sqrt{3}r)}{\sqrt{\left((2\sqrt{3})^2 + 6^2\right)}}, \quad (3)$$

$u3$ and $u2$ are two centroid points which are symmetric with respect to a given line segment BF. The equation of the line through points B and F is given by

$$\left(2\sqrt{3}\right)x - 6y + \sqrt{3}r = 0. \quad (4)$$

Thus,

$$u3 \cdot x = u2 \cdot x - \frac{2 * (2\sqrt{3}) * ((2\sqrt{3}) * u2 \cdot x - 6 * u2 \cdot y + \sqrt{3}r)}{\sqrt{\left((2\sqrt{3})^2 + 6^2\right)}},$$
$$u3 \cdot y = u2 \cdot y - \frac{2 * 6 * ((2\sqrt{3}) * u2 \cdot x - 6 * u2 \cdot y + \sqrt{3}r)}{\sqrt{\left((2\sqrt{3})^2 + 6^2\right)}}, \quad (5)$$

$u4$ and $u3$, $u5$ and $u2$, $u6$ and $u1$, these centroid points are symmetric with respect to y-axis, respectively.

Thus,

$$u4 \cdot x = -u3 \cdot x, \quad u4 \cdot y = u3 \cdot y,$$
$$u5 \cdot x = -u2 \cdot x, \quad u5 \cdot y = u2 \cdot y, \quad (6)$$
$$u6 \cdot x = -u1 \cdot x, \quad u6 \cdot y = u1 \cdot y.$$

The Lower triangle BDC and Upper triangle ABC are symmetric with respect to x-axis. The centroid of region in the Lower triangle can be obtained by the symmetric centroid in the Upper triangle.

Case 2 (4-connectivity). Any four points in 4-connectivity case form two equilateral triangles, where one is Upper triangle, and the other is Lower triangle, for example, equilateral Upper triangle *ABC* and Lower triangle *BDC* in *ABDC* 4-connectivity case of Figure 3. Without loss of generality, the Regions (11), (12), (15), and (16) in Figure 3 are in the Upper triangle, and the Regions (9), (10), (13), (14), (17), and (18) in Figure 3 are in the Lower triangle. We denote different restricted region in 4-connectivity case by U.7∼U.12 in the Upper triangle and L.7∼L.12 in the Lower triangle in Figure 4. If we can obtain the centroid of region U.7 in Figure 4, any other centroid of region will be determined by the geometrical symmetry.

The centroid of region U.7 denoted as $u7$ is obtained by

$$u7 \cdot x = \frac{\int_0^{r/2} \int_0^{\sqrt{r^2-x^2}-(\sqrt{3}/2)r} x \, dy \, dx}{\int_0^{r/2} \int_0^{\sqrt{r^2-x^2}-(\sqrt{3}/2)r} dy \, dx},$$
$$u7 \cdot y = \frac{\int_0^{r/2} \int_0^{\sqrt{r^2-x^2}-(\sqrt{3}/2)r} y \, dy \, dx}{\int_0^{r/2} \int_0^{\sqrt{r^2-x^2}-(\sqrt{3}/2)r} dy \, dx}. \quad (7)$$

Similar to the symmetric centroids obtained in 3-connecivity case, other centroids in 4-connectiviy case can be obtained easily.

Once we obtain the relative centroid of different restricted regions in Figure 4, any unknown node estimated absolute position will be calculated by coordinate shift. For example, we suppose that an unknown node q only contacts the mobile beacon three times. The three beacon positions are $(p1 \cdot x, p1 \cdot y)$, $(p2 \cdot x, p2 \cdot y)$, and $(p3 \cdot x, p3 \cdot y)$, where $p1 \cdot x > p2 \cdot x$; namely, these three beacon positions form an Upper triangle. If the relative position of unknown node belongs to Region U.1 in the Upper triangle, the estimated position q is calculated by

$$(q \cdot x, q \cdot y) = \left(u1 \cdot x + p1 \cdot x, u1 \cdot y + p1 \cdot y - \left(\frac{\sqrt{3}}{2} \right) r \right). \quad (8)$$

To avoid the integral operation of centroid calculation in the unknown node, the beacon transmits the calculated relative coordinates about $u1$ and $u7$ with specified radio range when contacting the unknown node for the first time. Then, any other centroid of region will be determined by the geometrical symmetry.

As described above, any unknown node only retains a limited number of RSS values (mainly 3 or 4) in the memory and requires simple arithmetic to solve the translation of position. Namely, MRC is a lightweight localization approach.

3.4. Irregular Radio Model. Our previously proposed localization approach MRC assumes an ideal radio scenario, where location sensory data are not influenced by the irregular radio range; that is, any receiver within the radio range of sender will hear the packets from that sender. However, on the one hand, variability in actual radio transmission patterns can have a substantial impact on localization accuracy depending on the localization technique [28]. On the other hand, the packet reception depends not only on the sender but also on the receiver. To this end, a perfect circular radio model assumption on real testbeds could be invalid for WSNs.

In this section, to ensure that MRC performs as true to reality as possible, we use a more general radio model [17] in our assumption. As the model described in [17], we assume an irregular radio transmission model with an upper and lower bounds on signal propagation. Namely, beyond the upper bound, all are out of communication range. Within the lower bound, every sensor node is guaranteed to be within communication range.

If the radio range is between these two boundaries, four scenarios are possible: (1) less than or equal to 2-connectivity, (2) 3-connectivity, (3) 4-connectivity, and (4) greater than or equal to 5-connecitvity.

Case 1 (less than or equal to 2-connectivity). If the radio transmission suffers severe interference effects from environmental noise, the unknown node may hear the connectivity observation only 1 or 2 times from the mobile beacon. As shown in Figures 5(a) and 5(b), the unknown node observes the connectivity from the beacon position in the solid black dot and does not hear from hollow black spot.

Figure 6 shows 1-connectivity case; the unknown node is out of any other beacon positions' radio range, such as *A*, *B*, *C*, *D*, *E*, and *F*, but within the radio range of the beacon position *O*. Then, the possible location of the unknown node is within the nonshadow coverage area, that is, near the beacon position *O* in Figure 6. Thus, the unknown node takes the only contacted beacon location *O* as its estimated location.

Case 2 (3-connectivity). Figures 5(c), 5(d), and 5(e) show all possible mutual position scenarios of three different beacon positions, respectively: (1) equilateral triangle, (2) isosceles triangle, and (3) collinear.

To distinguish three kinds of scenarios, the unknown node calculates the distance between any two beacon positions. Then, it calculates the number of those distances equal to the radio range. If the number is equal to three, the unknown node is within the equilateral triangle formed by three different beacon positions. We call this equilateral triangle form case regular 3-connectivity. If the number is equal to two, the node is within the isosceles triangle or near the line segment (collinear). We call these cases irregular 3-connectivity.

Case 3 (4-connectivity). Figures 5(f) and 5(g) show all possible mutual position scenarios of four different beacon positions: (1) diamond, and (2) nondiamond.

To distinguish these two kinds of scenarios, the node calculates the distance between any two connectivity positions. Then, it calculates the number of those distances equal to the radio range. If the number is equal to five, it shows that the unknown node is within the diamond formed by four different beacon positions. We call this diamond form case regular 4-connectivity. If the number is equal to three, we call the nondiamond form case irregular 4-connectivity.

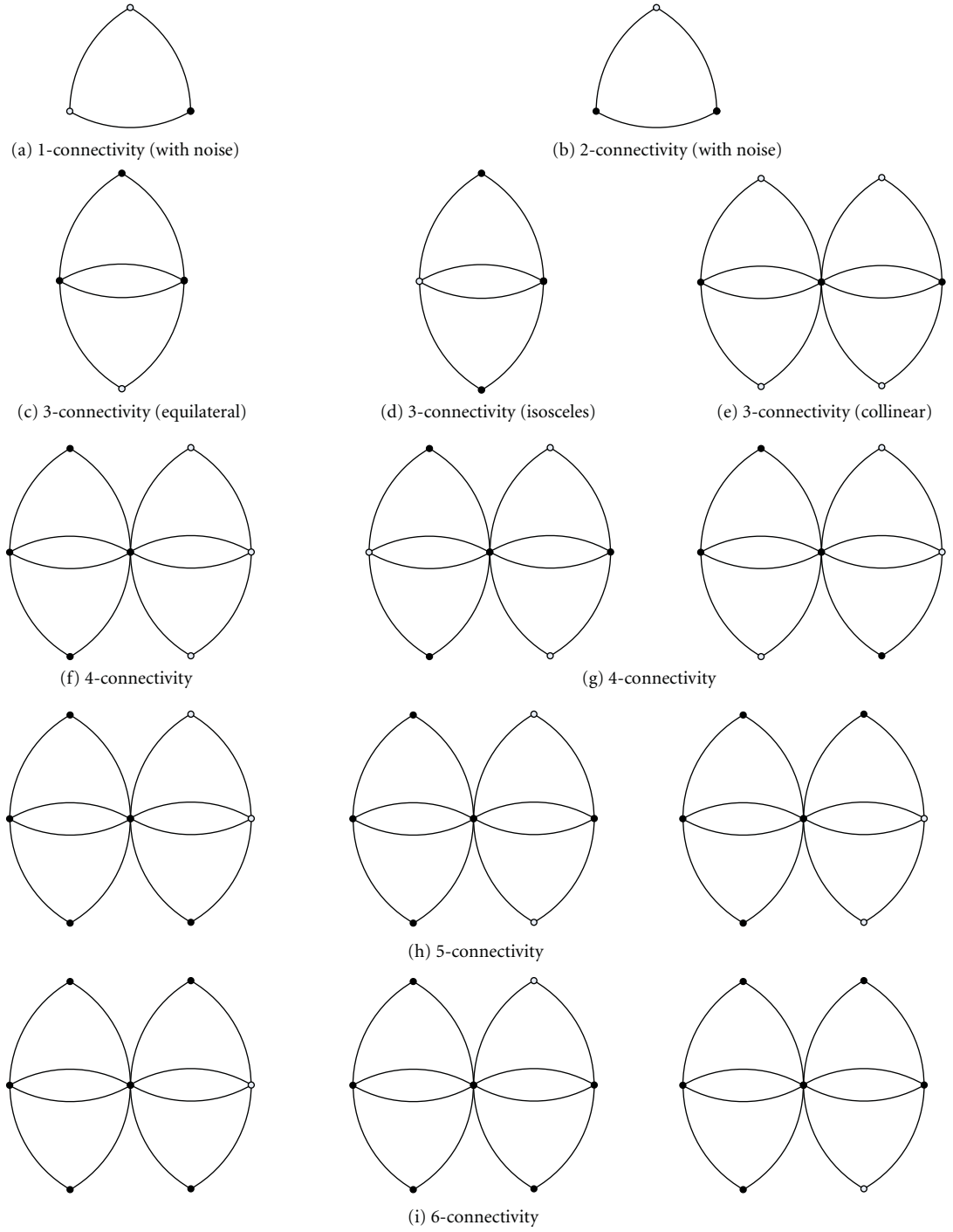


Figure 5: Connectivity of different cases.

Case 4 (greater than or equal to 5-connecitvity). As shown in Figures 5(h) and 5(i), the unknown node hears the beacon connectivity information greater than or equal to five times from the mobile beacon.

In general, these four scenarios can be divided into two types, regular connectivity and irregular connectivity. In the regular connectivity case, that is, regular 3-connectivity and regular 4-connecitvity, the unknown node will adopt 3-connectivity or 4-connectivity algorithm of MRC, respectively. While for the other case, that is, less than or equal to 2-connectivity, irregular 3-connectivity and 4-connectivity and greater than or equal to 5-connecitvity, we propose the following two solutions.

(i) The unknown node compares the RSS values obtained from those contacted beacon positions. Then, the node takes the beacon position making the unknown node receive the largest RSS value from

those beacon positions as the estimated location of the node. We call this method MRC_Nearest.

(ii) From the irregular beacon signals that it receives, the unknown node infers proximity to a collection of reference beacon positions. The unknown node localizes itself to the region which is defined by the centroid of these beacon positions. We call this solution MRC_Centroid.

Different from MRC, MRC_Nearest and MRC_Centroid can perform well in the irregular radio model by dealing with the above different connectivity cases between unknown node and beacon positions. In the evaluation section, we will compare the performance of the two solutions MRC_Nearest and MRC_Centroid in the noise environment.

Finally, let us consider 0-connectivity case; that is, the unknown nodes do not contact any beacon position. In the regular radio model, the mobile beacon with the trajectory will avoid 0-connectivity case. In the irregular radio model, 0-connectivity case is inevitable, but the probability is very small and has little impact on the result of average position accuracy. Once 0-connectivity of some unknown node appears, the unknown node will adopt some localization algorithms (e.g., Centroid localization algorithm) with localized neighbor sensor nodes after the mobile beacon has finished localization.

4. Simulation Evaluation

This section compares the performance with three typical range-free localization approaches closely related to our approaches: (1) classic range-free localization algorithms, (2) connectivity-based mobile-beacon-assisted range-free localization algorithms, and (3) RSS-assisted range-free localization algorithms.

The first type, such as Centroid, DV-hop, APIT, MDS-MAP, and SDP, is mainly designed for a static sensor network with multiple static anchors and unknown nodes. In order to obtain high redundancy of beacons without increasing deployment costs, these methods also can use a single moving anchor that sends out beacons at different locations to localize all unknown nodes inside the sensor network.

The second type, such as MSL and MBL, is specially designed for a static sensor network with a single mobile beacon and static unknown nodes. These mobile-beacon-assisted range-free approaches localize unknown nodes based on connectivity observation only.

The third type, such as PI, RSSL, and RSD, utilizes the variance of RSS to estimate the unknown node position, where the original design goal of PI and RSSL is to work in the mobile-assisted scenario, while RSD can capture the relative distance among 1-hop neighboring nodes to improve performance of the connectivity-base range-free approaches. Thus, RSD could augment some classic range-free localization algorithms, such as DV-hop, called DV-RSD, where the relative distance turns to the smallest accumulated RSD instead of shortest-path hops in DV-hop.

In the following content, assumptions are given in Section 4.1, and compared with three typical localization algorithms, the evaluation results are shown in Sections 4.2, 4.3, and 4.4, respectively.

4.1. Assumption. The key metric [26] for evaluating a localization algorithm is the accuracy of the location estimates or localization error. This is computed as follows:

$$\text{Error} = \frac{1}{M}\sum_{i=1}^{M}\|e_i - R_i\|, \tag{9}$$

$$M = \frac{n_d S}{\pi r^2}, \tag{10}$$

where M is the number of unknown nodes and can be obtained from (10), R_i denotes the real location of the ith unknown node, e_i denotes the location estimate of the ith unknown node, and $\|e_i - R_i\|$ denotes the distance between locations e_i and R_i. n_d in (10) denotes the unknown node density, the average number of nodes in one-hop transmission range; that is, n_d is the average number of neighbors to an unknown node. S denotes the size of the deployment area. The errors shown in the simulation results are in terms of the radio range; that is, the errors shown are computed by dividing the error in (9) by the radio range of sensor node.

Most parameter settings for our simulations are those used in [26, 28]. Our results were obtained using sensor nodes randomly distributed in a 500 units $\times$ 500 units square field, that is, $S = 500 \times 500$. In our experiments, we set ideal radio range $r = 100$, the unknown node density $n_d = 10$. We adopt Walking GPS architecture [32] for the beacon and unknown nodes which significantly reduce the size of the code and data memory used on the sensor node. Other simulation parameters of the unknown node are based on the MicaZ sensor node.

4.2. Compared with Classic Range-Free Localization Approaches. In these experiments, we simulate DV-hop with triangulation procedure to correct the estimated location. We modify APIT so that perfect point-in-triangulation (PIT) test is feasible, because the unknown node within the beacon triangles is easily determined with the trajectory proposed in Section 3.2. We simulate MDS-MAP with classical MDS algorithms. We simulate SDP in Matlab with the generic optimization solver SDPT3, called by CVX [33]. SDPT3 is a state-of-the-art primal-dual interior-point solver that exploits sparsity. Our experimental results are the average of 100 executions with different pseudorandom number generator seeds.

4.2.1. Localization Error When Varying Unknown Node Density

Random Deployment of Beacons. For the classical range-free localization algorithms, the deployment of beacons has no specific way. We put the same number of static anchors as the beacons in the MRC trajectory, randomly deployed in a two-dimensional plane. Due to the small number of random beacons, some unknown nodes will not be restricted

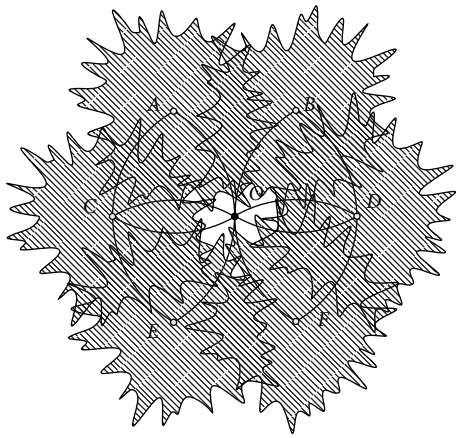


Figure 6: Connectivity with noise.

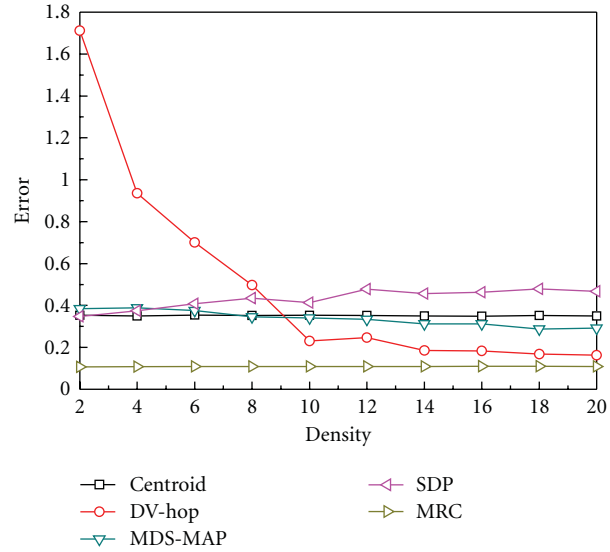


Figure 7: Error varying unknown node density with random deployment of beacons.

by enough beacons. Thus, APIT algorithm is invalid in this case. Figure 7 explores the effect of unknown node density on the localization estimation accuracy. Since, there is no interaction between unknown nodes in Centroid, and MRC algorithms, we see nearly constant results while varying unknown node density. For DV-hop, localization error decreases as the number of neighbors increases. As DV-hop, algorithms MDS-MAP and SDP rely on the observations from the neighbors including the unknown nodes and the beacons, but localization results have little change with different unknown node density. On the whole, our proposed method MRC improves the accuracy of nearly 30%–75% with the specific trajectory of mobile beacon in the perfect circular radio model.

Regular Deployment of Beacons. In this experiment, we compare these algorithms under a more consistent condition. Namely, we will deploy the anchors as the trajectory positions of MRC proposed in Section 3.2. Compared with Figure 7, the precision of Centroid and SDP algorithms have been improved nearly 50% due to more uniform anchors deployment, while the precision of DV-hop decreases due to more accurate 1-hop estimation in the random deployment scenario. Compared with the classic range-free localization algorithms with the same anchors positions as the trajectory beacons, the performance advantage of MRC has no significant change in Figure 8.

4.2.2. Localization Error When Varying DOI. We use degree of irregularity (DOI) to denote the maximum radio range variation in the direction of radio propagation. For example, if DOI = 0.1, then the actual radio range in each direction is randomly chosen from $[0.9r, 1.1r]$. In our experiments, the radio transmission range changes in time. SDP currently does not provide solutions in the noise environment.

Random Deployment of Beacons. As shown in Figure 9, the Centroid and MDS-MAP approaches are not substantially affected when DOI is between 0 and 0.5. The precision of DV-hop, MRC_Nearest, and MRC_Centroid algorithms is decreased as the value of DOI increased. MRC_Centroid method is always better than DV-hop and MRC_Nearest approaches, no matter the change in DOI.

Regular Deployment of Beacons. DV-hop does not work in this case. As shown in Figure 10, the MDS-MAP algorithm is not substantially affected. However, for Centroid, APIT, MRC_Nearest, and MRC_Centroid approaches, localization error increases as well as the value of DOI. Compared to classic range-free localization algorithms when varying DOI from 0 to 0.5, the accuracy of MRC_Centroid outperforms all the other approaches. As a result, in the irregular connectivity case, we conclude that the estimated position of unknown node is defined by the centroid of those contacted beacon positions.

4.3. Compared with Connectivity-Based Mobile-Beacon-Assisted Range-Free Localization Approaches. In this experiment, we compare MRC algorithms with some connectivity-based mobile-beacon-assisted range-free localization approaches, such as MSL, MBL. In general, as the time goes on, the mobile beacon contacts more unknown nodes; that is, more unknown node will be localized. The average localization accuracy of the mobile-beacon-assisted approaches will continue to be increased. Thus, the comparison of different localization convergences is much important. We simulate MSL and MBL with the number of particles for an unknown node $N = 50$ and adopt our proposed trajectory in Section 3.2 as MRC.

4.3.1. Localization Error As the Time Increased without Noise. Figure 11 shows the comparison of accuracy for MSL, MBL, and MRC under the same trajectory of the mobile beacon with the perfect circular radio model. In general, in all

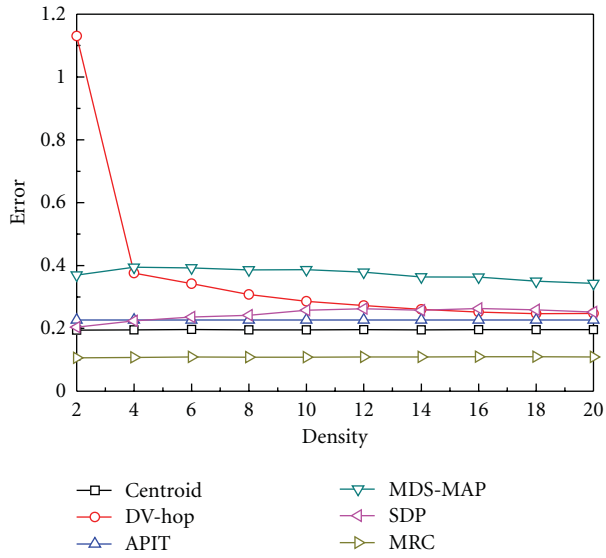


Figure 8: Error varying unknown node density with regular deployment of beacons.

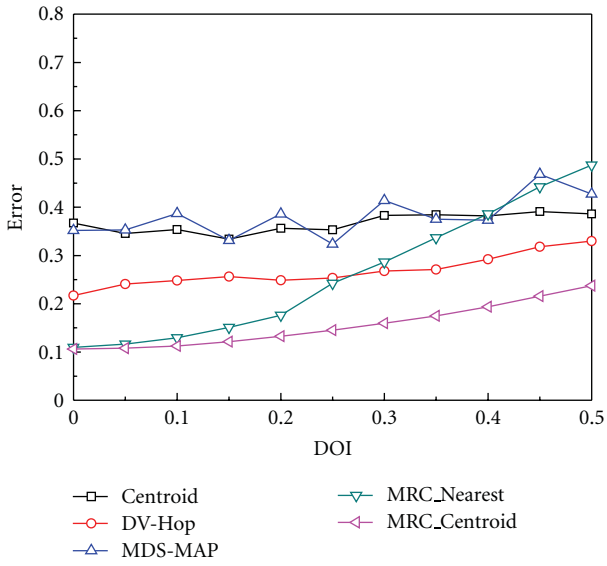


Figure 9: Error varying DOI with random deployment of beacons.

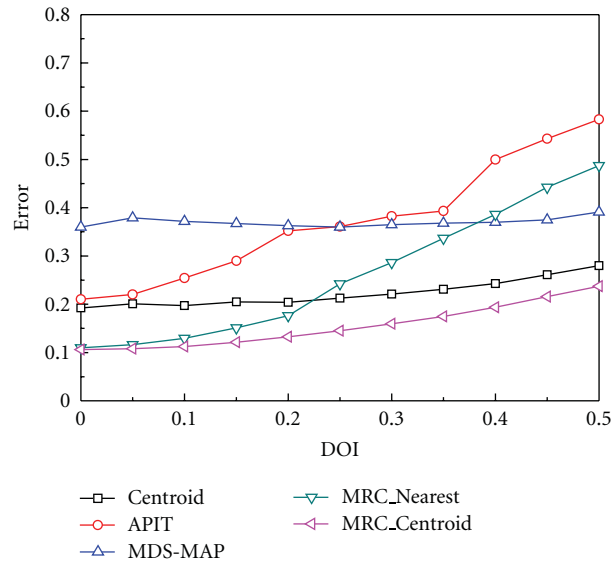


Figure 10: Error varying DOI with regular deployment of beacons.

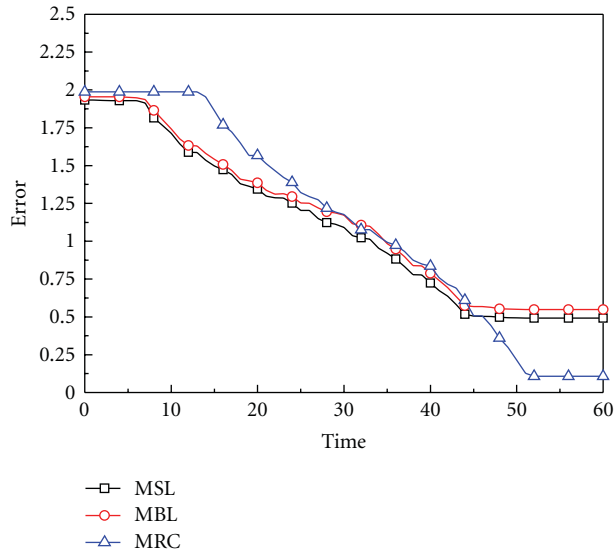


Figure 11: Compared with mobile-assisted localization without noise.

localization approaches, error decreases as the time increases. MSL and MBL utilize the mean of particles as the initial estimated position of unknown node. As shown in Figure 11, MSL and MBL, these two curvatures of the curves are very similar to each other. The results remaining unchanged denote the end of localization. When the accuracy of all approaches keeps stable, MRC improves nearly 80% compared with MSL and MBL.

4.3.2. Localization Error When Varying DOI. Figure 12 shows the comparison result of MSL, MBL, and MRC_Nearest and MRC_Centroid with the impact of DOI from 0 to 0.5. The localization error of all these approaches increases as the value of DOI. The cures of MSL and MBL are very close to each other. Compared to these two mobile-beacon-assisted range-free localization algorithms when varying DOI from 0 to 0.5, the accuracy of MRC_Centroid approach outperforms both of them.

4.4. Compared with RSS-Assisted Range-Free Localization Approaches. In this experiment, we adopt original RSS movement pattern [23] for RSSL methods. Suppose that the beacon is moving along the x-axis with broadcasting interval to be $r/2$ and the distance between two moving straight lines of the beacon is shortened (equal to r) to ensure that each unknown node can receive signals from the beacon. We simulate PI with original proposed optimal trajectory which consists of multiple joint optimal virtual triangles (VTs) and

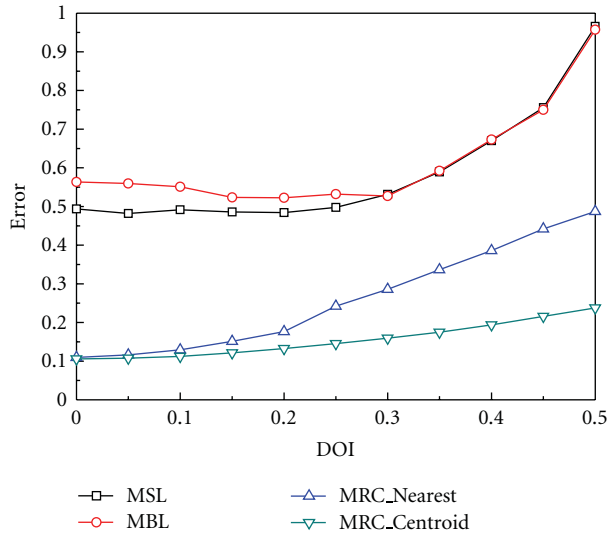


Figure 12: Compared with mobile-assisted localization with noise.

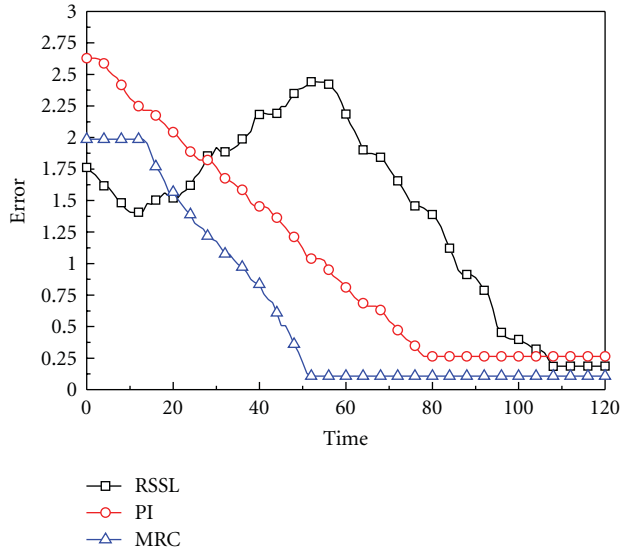


Figure 13: Compared with RSS-assisted localization without noise.

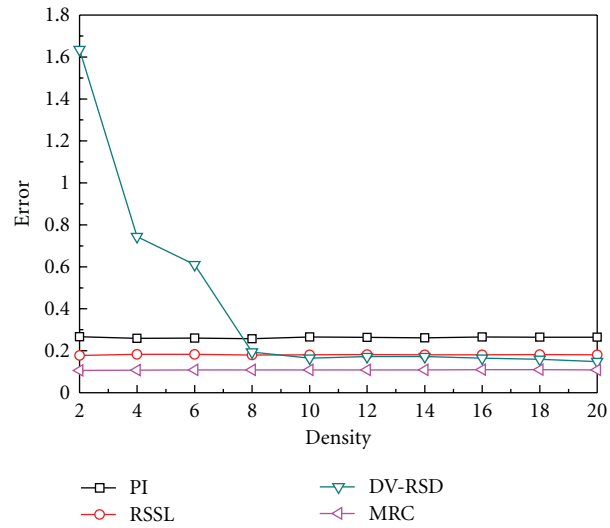


Figure 14: Compared with RSS-assisted localization varying unknown node density.

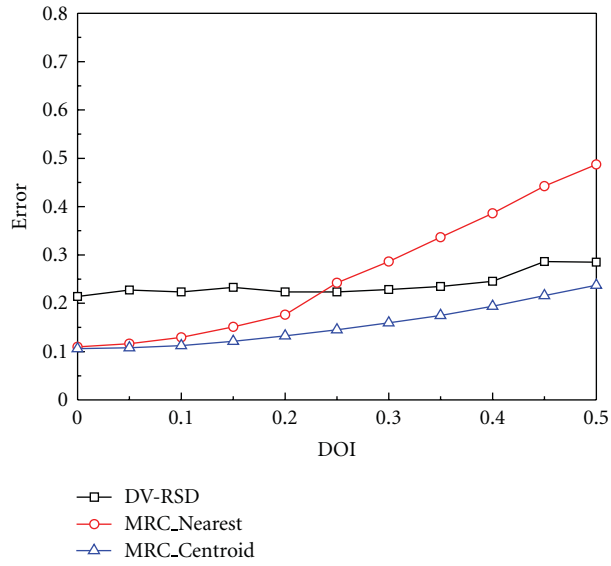


Figure 15: Compared with RSS-assisted localization with noise.

adopts random position in the deployment area as the initial estimated position of unknown node.

4.4.1. Localization Error As Beacons Increased without Noise. Considering the tradeoff between the trajectory length and the localization precision of RSS-based mobile-beacon-assisted approaches, we set the trajectory of each method to achieve the similar accuracy in the limited time. As shown in Figure 13, the trajectory of MRC spends the shortest time and obtains the highest accuracy than all the other approaches.

4.4.2. Localization Error When Varying Unknown Node Density. As the value of density increased, the curve of PI, RSSL, and MRC remains stable in Figure 14, because nonneighbor unknown node observation participated in localization process. The precision of MRC is always higher than the PI and RSSL. The estimation error of DV-RSD decreases as the number of neighbors increases. MRC is much better than DV-RSD in the low density of unknown nodes and still better in higher density, such as the density equal to 20.

4.4.3. Localization Error When Varying DOI. RSS-based approaches, PI and RSSL, currently do not provide solutions in the noise environment. Figure 15 shows that the overall location precision decreased as the value of DOI increased. MRC_Centroid always performs better than DV-RSD when DOI varies from 0 to 0.5.

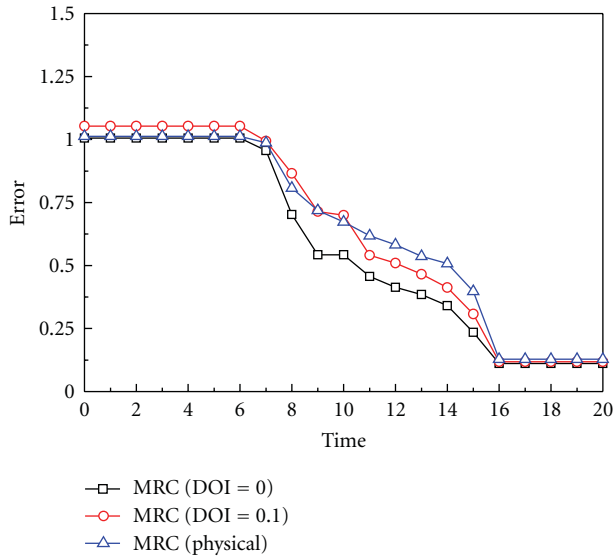


Figure 16: Compared with simulation and physical experiments.

5. Physical Evaluation

In this section, we give the physical evaluation of our localization algorithms in the outdoor environment. 50 MicaZ sensor nodes were randomly deployed 250 × 250 feet square area in the school playground. Each unknown node, including mobile beacon carried by the testers, was left 10 inches above the ground with antennas pointing to the sky. The radio sending power was 0 dBm at channel 26 (F_c = 2480 MHz) to avoid possible WiFi interference from the surroundings [2]. When the unknown node received a packet from the mobile beacon, it logged the sensed signal strength and the position of mobile beacon into its flash memory.

First, we measure the maximum distance (in feet) of 1-hop between unknown nodes in this special practical environment. The result of this measurement shows that 1-hop radio range varies from about 94 feet to 103 feet among different unknown node pairs along diverse directions. As shown in Section 3.2, the minimum distance between the transmitted beacon points is equal to 1-hop radio range. To this end, the distance between any two adjacent beacon positions of movement is 100 feet in this scenario, and DOI mentioned in Section 4.4 is less than 0.1.

Figure 16 shows the comparison of accuracy for MRC under simulation and physical environments. In general, in MRC within all different environments, error decreases as the time increases. The results of MRC (DOI = 0) and MRC (DOI = 0.1) are obtained in the simulation environment, and MRC (physical) is obtained in the outdoor environment. All these curvature of the curves are very similar to each other. The results remaining unchanged denote the end of localization. When the accuracy of all experiments keeps stable, results have no significant differences between simulation and physical environment; that is, MRC localization algorithm performs well in the actual environment.

6. Conclusion

We propose a distributed MRC localization approach with a specific trajectory in static WSNs. To ensure that MRC performs as true to reality as possible, we propose two improved approaches based on MRC to consider irregular radio scenario in the noisy environment. Evaluation shows that MRC_Centroid is more robust than MRC_Nearest. Comparing the performance with three typical range-free localization methods, our lightweight MRC algorithm with limited computation and storage overhead is more suitable for very low-computing power sensor nodes that can efficiently outperform better with only basic arithmetic operations in the ideal or irregular radio model. As future work, more practical factors will be considered, and different practical factors may affect efficiency, accuracy, and flexibility of our localization approaches.

Acknowledgments

This work is supported by the Key Science and Technology Innovation Team Constructive Projects of Zhejiang province, China under Grant no. 2009R50009 and also supported by the National Basic Research Program of China (973 Program) under Grant no. 2006CB303000.

References

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] Z. Zhong and T. He, "Achieving range-free localization beyond connectivity," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, (SenSys '09)*, pp. 281–294, Berkeley, Calif, USA, November 2009.

[3] G. Zhongwen, G. Ying, H. Feng et al., "Perpendicular intersection: locating wireless sensors with mobile beacon," in *Proceedings of the Real-Time Systems Symposium, (RTSS '08)*, pp. 93–102, Barcelona, Spain, December 2008.

[4] B. Hofmann-Wellenhof, H. Lichtenegger,, and J. Collins, *Global Positioning System: Theory and Practice*, Springer, New York, NY, USA, 4th edition, 1997.

[5] S. Lanzisera, D. T. Lin, and K. S. J. Pister, "RF time of flight ranging for wireless sensor network localization," in *Proceedings of the 4th Workshop on Intelligent Solutions in Embedded Systems, (WISES '06)*, pp. 165–176, Vienna, Austria, June 2006.

[6] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "Cricket location-support system," in *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MOBICOM '00)*, pp. 32–43, Boston, Mass, USA, August 2000.

[7] A. Savvides, C. C. Han, and M. B. Strivastava, "Dynamic fine-grained localization in ad-hoc networks of sensors," in *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, pp. 166–179, Rome, Italy, July 2001.

[8] X. Cheng, A. Thaeler, G. Xue, and D. Chen, "TPS: a time-based positioning scheme for outdoor wireless sensor networks," in *Proceedings of the 23rd Annual Joint Conference of the IEEE*

Computer and Communications Societies, vol. 4, pp. 2685–2696, Hong Kong, 2004.

[9] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," in *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications*, vol. 3, pp. 1734–1743, San Francisco, Calif, USA, 2003.

[10] H. Chang, J. Tian, T. Lai, H. Chu, and P. Huang, "Spinning beacons for precise indoor localization," in *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, pp. 127–140, Raleigh, NC, USA, 2008.

[11] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communications Societie*, pp. 775–784, Tel Aviv, Israel, 2000.

[12] J. A. Costa, N. Patwari, and A. O. Hero, "Distributed weighted-multidimensional scaling for node localization in sensor networks," *ACM Transactions on Sensor Networks*, vol. 2, no. 1, pp. 39–64, 2006.

[13] A. Varshavsky, E. de Lara, J. Hightower, A. LaMarca, and V. Otsason, "GSM indoor localization," *Pervasive and Mobile Computing*, vol. 3, no. 6, pp. 698–720, 2007.

[14] K. Yedavalli and B. Krishnamachari, "Sequence-based localization in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 1, pp. 81–94, 2008.

[15] Y. Zhang, L. Zhang, and X. Shan, "Ranking-Based statistical localization for wireless sensor networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference, (WCNC '08)*, pp. 2561–2566, Las Vegas, Nev, USA, April 2008.

[16] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," *IEEE Personal Communications*, vol. 7, no. 5, pp. 28–34, 2000.

[17] T. He, C. Huang, B. M. Blum, J. A. Stankovic, and T. Abdelzaher, "Range-free localization schemes for large scale sensor networks," in *Proceedings of the 9th Annual International Conference on Mobile Computing and Networking, (MobiCom '03)*, pp. 81–95, San Diego, Calif, USA, September 2003.

[18] D. Niculescu and B. Nath, "DV based positioning in Ad Hoc networks," *Telecommunication Systems*, vol. 22, no. 1-4, pp. 267–280, 2003.

[19] Y. Shang and W. Ruml, "Improved MDS-based localization," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 4, pp. 2640–2651, Hong Kong, 2004.

[20] Q. Shi and C. He, "A SDP approach for range-free localization in wireless sensor networks," in *Proceedings of the IEEE International Conference on Communications, (ICC '08)*, pp. 4214–4218, Beijing, China, May 2008.

[21] K. Römer, "The Lighthouse Location System for Smart Dust," in *Proceedings of the 1st International Conference on Mobile systems, Applications and Services*, pp. 15–30, San Francisco, Calif, USA, 2003.

[22] R. Stoleru, T. He, J. A. Stankovic, and D. Luebke, "A High-accuracy, low-cost localization system for wireless sensor networks," in *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, pp. 13–26, San Diego, Calif, USA, 2005.

[23] B. Xiao, H. Chen, and S. Zhou, "A walking beacon-assisted localization in wieless sensor networks," in *Proceedings of the IEEE International Conference on Communications*, pp. 3070–3075, Glasgow, Scotland, 2007.

[24] G. Teng, K. Zheng, and W. Dong, "An efficient and self-adapting localization in static wireless sensor networks," vol. 9, no. 8, pp. 6150–6170, 2009.

[25] R. Huang and G. V. Záruba, "Static path planning for mobile beacons to localize sensor networks," in *Proceedings of the 5th Annual IEEE International Conference on Pervasive Computing and Communications Workshops, (PerCom '07)*, pp. 323–328, New York, NY, USA, March 2007.

[26] M. Rudafshani and S. Datta, "Localization in wireless sensor networks," in *Proceedings of the 6th International Symposium on Information Processing in Sensor Networks (IPSN '07)*, pp. 51–60, Cambridge, Mass, USA, April 2007.

[27] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[28] L. Hu and D. Evans, "Localization for mobile sensor networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pp. 45–57, New York, NY, USA, October 2004.

[29] M. L. Sichitiu and V. Ramadurai, "Localization of wireless sensor networks with a mobile beacon," in *Proceedings of the IEEE International Conference on Mobile Ad-Hoc and Sensor Systems*, pp. 174–183, Philadelphia, Pa, USA, October 2004.

[30] N. B. Priyantha, H. Balakrishnan, E. D. Demaine, and S. Teller, "Mobile-assisted localization in wireless sensor networks," in *Proceedings of The 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 1, pp. 172–183, Miami, Fla, USA, 2005.

[31] K. Kim and W. Lee, "MBAL: a mobile beacon-assisted localization scheme for wireless sensor networks," in *Proceedings of the 16th International Conference on Computer Communications and Networks 2007, (ICCCN '07)*, pp. 57–62, Honolulu, Hawaii, USA, August 2007.

[32] R. Stoleru, T. He, and J. A. Stankovic, "Walking GPS: a practical solution for localization in manually deployed wireless sensor networks," in *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, (LCN '04)*, pp. 480–489, Tampa, Fla, USA, November 2004.

[33] M. Grant and S. Boyd, "CVX: matlab software for disciplined convex programming," http://cvxr.com/cvx/.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 172902, 8 pages
doi:10.1155/2011/172902

Research Article

Total Least Squares Method for Robust Source Localization in Sensor Networks Using TDOA Measurements

Yang Weng,[1] Wendong Xiao,[2] and Lihua Xie[3]

[1] *College of Mathematics, Sichuan University, Chengdu 610064, China*
[2] *Networking Protocols Department, Institute for Infocomm Research, Singapore 138632*
[3] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798*

Correspondence should be addressed to Wendong Xiao, wxiao@i2r.a-star.edu.sg

Received 31 March 2011; Accepted 15 June 2011

The TDOA-based source localization problem in sensor networks is considered with sensor node location uncertainty. A total least squares (TLSs) algorithm is developed by a linear closed-form solution for this problem, and the uncertainty of the sensor location is formulated as a perturbation. The sensitivity of the TLS solution is also analyzed. Simulation results show its improved performance against the classic least squares approaches.

1. Introduction

Driven by many practical applications such as environment monitoring, traffic management in intelligent transportation, healthcare for the older and disabled, cyber-physical system (CPS) has emerged as a new advanced system to link the virtual cyber world with the real physical world [1, 2]. Sensor network is crucial to enable CPS by efficient monitoring and understanding the physical world. The main purpose of a sensor network is to monitor an area, including detecting, identifying, localizing, and tracking one or more objects of interest. These networks may be used by the military in surveillance, reconnaissance, and combat scenarios or around the perimeter of a manufacturing plant for intrusion detection. The problem of source localization involves the estimation of the position of a stationary transmitter from multiple noisy sensor measurements, which can be time-of-arrival (TOA), time-difference-of-arrival (TDOA), or angle-of-arrival (AOA) measurements or a combination of them [3–5].

Source localization methods using TDOA measurements locate the source at the intersection of a set of hyperboloids. Finding this intersection is a highly nonlinear problem [6, 7]. Over the years, many iterative numerical algorithms have been proposed for the problem, including the maximum likelihood estimation methods [8, 9] and the constrained optimization methods [10, 11]. In these approaches, linear approximation and iterative numerical techniques have to be used to deal with the nonlinearity of the hyperbolic equations. However, it is difficult to select a good initial guess to avoid a local minimum for them; therefore the convergence to the optimal solution cannot be guaranteed. The closed-form solution methods are widely used since no initial solution guesses are required and have no divergence problem compared with the iterative techniques [12–16]. For real-time application in WSNs, the iterative procedure for iterative algorithm is time consuming while the closed-form method is computational efficient.

The aforementioned approaches need the precise location of sensors. In practice, the receiver locations may not be known exactly. For example, in sensor network applications, the receivers can be with airplanes or unmanned aerial vehicles (UAVs) whose positions and velocities may not be precisely known. Hence, the inaccuracy in receiver locations needs to be taken into account in practical applications which is challenging and difficult as the estimation performance of source location can be very sensitive to the accurate knowledge of the receiver positions and a slight error in a receiver's location can lead to a big error in the source location estimate. In [17], a closed-form solution is proposed that takes the receiver error into account to reduce the estimation error. The proposed solution is computationally efficient and does not have the divergence problem as in the iterative techniques. In [18], the maximum likelihood formulation

of source localization problem is given and an efficient convex relaxation for this nonconvex optimization problem is proposed. A formulation for robust source localization in the presence of sensor location errors is also proposed. Both the above methods assume that the measurement noise is Gaussian and characterize the uncertainty by stochastic approach with the perturbation being white and Gaussian.

However, such white and Gaussian assumptions are unrealistic in many practical applications [19–21]. Usually, if the Gaussian assumption are not met, the maximum-likelihood-based results under Gaussian assumption may lead to poor estimation performance, which means that the estimation performance is sensitive to the exact knowledge of the parameters of the system (see, e.g., [22]). These facts motivate us to further research on robust source localization method without any distribution assumption for measurements noise and sensor location error.

In this paper, we will develop a total least squares (TLSs) algorithm for location estimation of a stationary source. TLS is a least squares data modeling technique in which observational errors on both dependent and independent variables are taken into account. The uncertainty of the sensor location is formulated as a perturbation on the given sensor location. The sensitivity of the TLS solution will also be analyzed to show the superiority of our proposed algorithm. Compared with the existing methods which need the Gaussian assumption for both measurements noise and sensor location error, the TLS approach does not depend on any assumed distribution of the noise and errors. Simulation results support the above analysis and show good performance of the proposed method.

The rest of this paper is organized as follows. In Section 2, a linear closed-form solution is given for source localization problem using TDOA measurements. The total least squares method for source localization with sensor location uncertainty is given in Section 3, and the corresponding sensitivity analysis is derived in Section 4. Simulation results are presented in Section 5 to show the improved performance of the proposed method against the classic least squares approaches. Concluding remarks are made in Section 6.

2. A Linear Closed-Form Solution

Assume that sensor i is located at point $S_i = \{x_i, y_i, z_i\}$. Denote the unknown source location by $S = \{x, y, z\}$. Let c be the signal propagation speed and N the number of sensor nodes distributed in the network. A reference node, denoted as S_0, exists in the field. We define the distance from the source to sensor i as D_i. The TDOA that we derive from the data, τ_{0i}, for each sensor node $i \in [1, N]$ relative to the reference sensor S_0 is

$$\tau_{0i} = \frac{1}{c}\|S - S_0\| - \frac{1}{c}\|S - S_i\| = \frac{1}{c}(D_0 - D_i). \tag{1}$$

Denote the distance difference of arrival (DDOA) data for the ith sensor as

$$d_{0i} = \tau_{0i} \times c = D_0 - D_i. \tag{2}$$

We have

$$\begin{aligned} D_0^2 - D_i^2 &= \|S - S_0\|^2 - \|S - S_i\|^2, \\ &= (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2, \\ &\quad - \left((x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2\right), \\ &= x_0^2 - x_i^2 + 2x(x_i - x_0) + y_0^2 - y_i^2 + 2y(y_i - y_0), \\ &\quad + z_0^2 - z_i^2 + 2z(z_i - z_0), \\ &= D_0^2 - (D_0 - d_{0i})^2, \\ &= 2D_0 d_{0i} - d_{0i}^2. \end{aligned} \tag{3}$$

Group all the known terms together and denote

$$b_i = \frac{1}{2}[x_0^2 - x_i^2 + y_0^2 - y_i^2 + z_0^2 - z_i^2 + d_{0i}^2]. \tag{4}$$

Rearranging and substituting give

$$(x_0 - x_i)x + (y_0 - y_i)y + (z_0 - z_i)z + d_{0i}D_0 = b_i, \tag{5}$$

which is a linear model for unknown parameters x, y, z, and D_0. Stacking the N sensor measurements, we have the linear system in matrix form

$$AX = b, \tag{6}$$

where

$$\begin{aligned} A &= \begin{bmatrix} x_0 - x_1 & y_0 - y_1 & z_0 - z_1 & d_{01} \\ x_0 - x_2 & y_0 - y_2 & z_0 - z_2 & d_{02} \\ \vdots & \vdots & \vdots & \vdots \\ x_0 - x_N & y_0 - y_N & z_0 - z_N & d_{0N} \end{bmatrix}, \\ X &= \begin{bmatrix} x \\ y \\ z \\ D_0 \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}. \end{aligned} \tag{7}$$

When the location of sensor nodes can be precisely known and the TDOA measurements are noise-free, linear system (6) is compatible. The solution is unique while the data matrix A is of full rank.

In [16], a noise ε is considered at the right side of (6):

$$AX = b + \varepsilon. \tag{8}$$

The least squares (LSs) problem seeks to

$$\begin{aligned} &\text{minimize :} \quad \|b - b_{\text{ls}}\|, \\ &\text{subject to :} \quad b_{\text{ls}} \in \mathcal{R}(A), \end{aligned} \tag{9}$$

where $\mathcal{R}(A)$ denote the vector space spanned by the column vector of matrix A. Once a minimizing b_{ls} is found, then any

X satisfying $AX = b_{\text{ls}}$ is called an LS solution and $b - b_{\text{ls}}$ the corresponding LS correction. The unique LS solution can be obtained while the data matrix A is of full rank:

$$X_{\text{ls}} = \left(A^T A\right)^{-1} A^T b. \tag{10}$$

The ordinary LS problem amounts to perturbing the observation vector b by a minimum amount $b - AX_{ls}$. The underlying assumption is that errors only occur in the vector b and that the matrix A is exactly known. However, the TDOA measurements always have noise, that is, d_{0i} is perturbed by a noise. Hence, both sides of (5) have perturbation according to term d_{0i}. Besides, in practice, the receiver locations may not be known exactly. As a result, the inaccuracy in receiver locations needs to be taken into account in practical environments. Therefore, considering only the perturbation at the right side of (6) is not enough. We should consider the perturbations at both sides.

3. Total Least Squares Solution

The definition of the total least squares method is motivated by the asymmetry of the least squares method that b is corrected while A is not. Provided that both A and b are given data, it is reasonable to treat them symmetrically. One important application of TLS problem is parameter estimation in errors-in-variables models, that is, considering the measurements in A [23, 24]. We assume that the m measurement in $\widetilde{A}$ and $\widetilde{b}$ by

$$\widetilde{A}X = \widetilde{b}, \tag{11}$$

where $\widetilde{A} = A + \Delta A$, $\widetilde{b} = b + \Delta b$, and $AX = b$, is compatible. The total least squares (TLSs) problem seeks to

$$\begin{aligned} &\text{minimize :} \quad \left\|[\widetilde{A}; \widetilde{b}] - [A_{\text{tls}}; b_{\text{tls}}]\right\|_F, \\ &\text{subject to :} \quad b_{\text{tls}} \in \mathcal{R}(A_{\text{tls}}), \end{aligned} \tag{12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of matrix A, that is,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j}^{n}\left|a_{ij}\right|^2} = \sqrt{\text{trace}(A^T A)}. \tag{13}$$

Once a minimizing $[A_{\text{tls}}; b_{\text{tls}}]$ is found, then any X satisfying $A_{\text{tls}}X = b_{\text{tls}}$ is called a TLS solution and $[\widetilde{A}; \widetilde{b}] - [A_{\text{tls}}; b_{\text{tls}}]$ the corresponding TLS correction [25].

When A is of full rank, the closed-form expression of the basic TLS solution can be obtained as the following:

$$X_{\text{tls}} = \left(\widetilde{A}^T\widetilde{A} - \sigma_{n+1}^2 I\right)^{-1}\widetilde{A}^T\widetilde{b}, \tag{14}$$

where σ_{n+1} is the smallest singular value of $[A; b]$. It can be proved that the TLS solution X_{tls} estimates the true parameter A^+b consistently, that is, X_{tls} converges to the solution of $AX = b$ as the number of measurements tends to infinity, where A^+ denotes the Moore-Penrose pseudoinverse of matrix A. This property of TLS estimates does not depend on any assumed distribution of the errors. Note that the LS estimates are inconsistent in this case.

In the following, the algorithm computation complexity is analyzed by considering the number of floating-point operations (FLOPS). The calculation of FLOPS is briefly described as follows: additions and multiplications count as one FLOP each. Adding matrices of sizes $m \times n$ requires mn FLOPS. Multiplying matrices of sizes $m \times k$ and $k \times p$ requires mnk FLOPS. Matrix inverse of size $n \times n$ requires n^3 FLOPS. Singular value decomposition of matrix with size $n \times m$ requires nm^2 FLOPS. The computation overhead of three closed-form algorithms is investigated, that is, least squares (LSs) method in [16], two-stage weighted least squares (WLSs) method in [17] and the proposed TLS method. For N-deployed nodes and p dimension of source location parameter, the numbers of FLOPS in the LS, WLS, and the proposed TLS algorithms are $(2p + 3)(p + 1)N + (p + 1)^3$, $(2p + 3)(2p + 2)N + 2(p + 1)^3$, and $(2p + 5)(p + 2)N + (p + 2)^2(p + 3)$, respectively. It is to show that LS method is the most computationally efficient, whereas WLS method is two-stage least square and TLS method has singular value decomposition for matrix. It is clear that they all have comparable computation complexity $\mathcal{O}(N)$ since $p = 3$ or $p = 4$ for source localization problem.

4. Sensitivity Properties of the Solution

In this section, we will first examine how perturbations in A and b affect the solution X. In this analysis the condition number of the matrix A plays a significant role. The following definition generalizes the condition number of a square nonsingular matrix [26].

Definition 1. Let $A \in \mathcal{R}^{m\times n}$ have rank r. The condition number of A is

$$\kappa(A) = \|A\| \cdot \left\|A^+\right\| = \frac{\sigma_1}{\sigma_r}, \tag{15}$$

where $\sigma_1, \sigma_2, \ldots, \sigma_r$ are the singular values of A by decreasing order.

Matrices with small condition numbers are said to be well conditioned while the ones with large condition numbers are said to be ill-conditioned. Analyzing the effect of perturbation on the solution of linear system $AX = b$, we introduce the following lemma [27].

Lemma 1. *If* $\text{rank}(\widetilde{A}) = \text{rank}(A)$ *and* $\|\Delta A\|\|A^+\| < 1$, *then*

$$\left\|\widetilde{A}^+\right\| \le \frac{\|A^+\|}{1 - \|A^+\|\|\Delta A\|}. \tag{16}$$

Theorem 1. *Denote* $X_{\text{ls}} = A^+b$ *as the LS solution of unperturbed system* $AX = b$, *and* $\widetilde{X}_{\text{ls}} = \widetilde{A}^+\widetilde{b}$ *is the LS solution of perturbed system* $\widetilde{A}X = \widetilde{b}$ *with* $\widetilde{A} = A + \Delta A$, $\widetilde{b} = b + \Delta b$. *One further assumes that* $\|\Delta A\| \le \varepsilon\|A\|$, $\|\Delta b\| \le \varepsilon\|b\|$, $\kappa(A) = \kappa$. *Then, one has*

$$\frac{\left\|\widetilde{X}_{\text{ls}} - X_{\text{ls}}\right\|}{\|X_{\text{ls}}\|} = \mathcal{O}(\varepsilon\kappa). \tag{17}$$

Proof. Noting that A and $\widetilde{A}$ are full rank and $(I - AA^{+})b = 0$ since $b \in \mathcal{R}(A)$, we have

$$\begin{aligned}
\widetilde{X}_{\mathrm{ls}} - X_{\mathrm{ls}} &= \widetilde{A}^{+}\widetilde{b} - A^{+}b \\
&= \widetilde{A}^{+}(b + \Delta b) - A^{+}b \\
&= \widetilde{A}^{+}\Delta b + \left(\widetilde{A}^{+} - A^{+}\right)b \\
&= \widetilde{A}^{+}\Delta b + \left(\widetilde{A}^{+} + \widetilde{A}^{+}AA^{+} - \widetilde{A}^{+}AA^{+} - A^{+}\right)b \\
&= \widetilde{A}^{+}\Delta b + \left(\widetilde{A}^{+} + \widetilde{A}^{+}AA^{+} - \widetilde{A}^{+}AA^{+} - \widetilde{A}^{+}\widetilde{A}A^{+}\right)b \\
&= \widetilde{A}^{+}\Delta b + \widetilde{A}^{+}(I - AA^{+})b - \widetilde{A}^{+}\left(\widetilde{A} - A\right)A^{+}b \\
&= \widetilde{A}^{+}\Delta b - \widetilde{A}^{+}\Delta A X_{\mathrm{ls}}.
\end{aligned} \tag{18}$$

Therefore, $\|\widetilde{X}_{\mathrm{ls}} - X_{\mathrm{ls}}\| = \|\widetilde{A}^{+}\Delta b - \widetilde{A}^{+}\Delta A X_{\mathrm{ls}}\|$. Dividing by $\|X_{\mathrm{ls}}\|$ at both sides, we have

$$\begin{aligned}
\frac{\left\|\widetilde{X}_{\mathrm{ls}} - X_{\mathrm{ls}}\right\|}{\|X_{\mathrm{ls}}\|} &= \frac{\left\|\widetilde{A}^{+}\Delta b - \widetilde{A}^{+}\Delta A X_{\mathrm{ls}}\right\|}{\|X_{\mathrm{ls}}\|} \\
&\le \frac{\widetilde{A}^{+}\Delta b}{\|X_{\mathrm{ls}}\|} + \left\|\widetilde{A}^{+}\Delta A\right\| \\
&\le \left\|\widetilde{A}^{+}\right\|\left(\frac{\|\Delta b\|}{\|X_{\mathrm{ls}}\|} + \|\Delta A\|\right) \\
&\le \left\|\widetilde{A}^{+}\right\|\left(\frac{\varepsilon\|b\|}{\|X_{\mathrm{ls}}\|} + \varepsilon\|A\|\right) \\
&\le \frac{\|A^{+}\|}{1 - \|A^{+}\|\|\Delta A\|}\left(\frac{\varepsilon\|b\|}{\|X_{\mathrm{ls}}\|} + \varepsilon\|A\|\right) \\
&\le \frac{\|A^{+}\|}{1 - \|A^{+}\|\varepsilon\|A\|}\left(\frac{\varepsilon\|b\|}{\|X_{\mathrm{ls}}\|} + \varepsilon\|A\|\right) \\
&= \frac{1}{1 - \varepsilon\kappa}\left(\frac{\varepsilon\|A^{+}\|\|b\|}{\|A^{+}b\|} + \varepsilon\kappa\right) \\
&= \frac{1}{1 - \varepsilon\kappa}\left(\frac{\varepsilon\|A^{+}\|\|AA^{+}b\|}{\|A^{+}b\|} + \varepsilon\kappa\right) \\
&\le \frac{1}{1 - \varepsilon\kappa}\left(\frac{\varepsilon\|A^{+}\|\|A\|\|A^{+}b\|}{\|A^{+}b\|} + \varepsilon\kappa\right) \\
&= \frac{2\varepsilon\kappa}{1 - \varepsilon\kappa} = \mathcal{O}(\varepsilon\kappa).
\end{aligned} \tag{19}$$

□

From the theorem we can see that, only when the perturbations are sufficiently small, the LS solution is a good estimator of the true solution of $AX = b$.

Additionally, we will show when the TLS solution has better performance than the LS solution for the perturbed model $\widetilde{A}X \approx \widetilde{b}$. Denote the singular value decomposition (SVD) of A by

$$A = U\Sigma V^{T}, \tag{20}$$

where

$$\begin{aligned}
&U = [u_1, u_2', \ldots, u_m], \quad U' \in \mathcal{R}^{m},\ U^{T}U = I_m, \\
&V = [v_1, v_2', \ldots, v_n], \quad V' \in \mathcal{R}^{n},\ V^{T}\mathrm{V} = I_n, \\
&\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathcal{R}^{m\times n}, \quad \sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0,
\end{aligned} \tag{21}$$

denote the singular value decomposition (SVD) of $\widetilde{A}$ by

$$A + \Delta A = \widetilde{A} = U'\Sigma' V'^{T}, \tag{22}$$

where

$$\begin{aligned}
&U' = [u_1', u_2', \ldots, u_m'], \quad U' \in \mathcal{R}^{m},\ U'^{T}U' = I_m, \\
&V' = [v_1', v_2', \ldots, v_n'], \quad V' \in \mathcal{R}^{n},\ V'^{T}V' = I_n, \\
&\Sigma' = \mathrm{diag}(\sigma_1', \sigma_2', \ldots, \sigma_n') \in \mathcal{R}^{m\times n}, \\
&\qquad \sigma_1' \ge \sigma_2' \ge \ldots \ge \sigma_n' \ge 0,
\end{aligned} \tag{23}$$

and denote the SVD of $[\widetilde{A}; \widetilde{b}]$ by

$$\left[A + \Delta A; b + \Delta b, \widetilde{A}; \widetilde{b}\right] = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{T}, \tag{24}$$

where

$$\begin{aligned}
&\widetilde{U} = [\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_m], \quad \widetilde{U} \in \mathcal{R}^{m},\ \widetilde{U}^{T}\widetilde{U} = I_m, \\
&\widetilde{V} = [\widetilde{v}_1, \widetilde{v}_2, \ldots, \widetilde{v}_n], \quad \widetilde{V} \in \mathcal{R}^{(n+1)},\ \widetilde{V}^{T}\widetilde{V} = I_{n+1}, \\
&\widetilde{\Sigma} = \mathrm{diag}(\widetilde{\sigma}_1, \widetilde{\sigma}_2, \ldots, \widetilde{\sigma}_n, \widetilde{\sigma}_{n+1}) \in \mathcal{R}^{m\times n+1}, \\
&\qquad \widetilde{\sigma}_1 \ge \widetilde{\sigma}_2 \ge \cdots \ge \widetilde{\sigma}_{(n+1)} \ge 0.
\end{aligned} \tag{25}$$

The accuracy of TLS and LS solutions related to the perturbation effects on singular values and associated singular subspaces [23]. Several papers have analyzed the bounds on the perturbation effects related to singular subspaces [28–30]. The most interesting results for sensitivity analysis of TLS solution are given in [30].

Definition 2. Denote two subspaces as L and M. For any unitary invariant norm, the distance between two subspaces is defined as the sine of the largest canonical angle

$$\mathrm{dist}(L, M) = \|\sin\Theta(L, M)\| = \|(I - P_M)P_L\|, \tag{26}$$

where P_M and P_L are the projection operators [29, 30].

Theorem 2. *Let the SVD of $A \in \mathcal{R}^{m\times n}, m \ge n$, be given by* (20), $\mathrm{rank}(A) = r$. *Add perturbations ΔA to A, and let SVD of $\widetilde{A}$ be given by* (22). *If $\sigma_r' - \sigma_{r+1} > 0$, then*

$$\mathrm{dist}\left(\mathcal{R}(A), \mathcal{R}\left(\widetilde{A}\right)\right) \le \frac{\|\Delta A\|}{\sigma_r' - \sigma_{r+1}}. \tag{27}$$

Theorem 2 is a special case of the generalized $\sin\theta$-theorem in [30].

Recall the perturbed system $\widetilde{A}X \approx \widetilde{b}$; LS projects $\widetilde{A}$ and $\widetilde{b}$ into the singular subspaces

$$\mathcal{L}(u_1', u_2', \ldots, u_m') = \mathcal{R}\left(\widetilde{A}\right), \tag{28}$$

while the TLS projects $\widetilde{A}$ and $\widetilde{b}$ into the singular subspaces

$$\mathcal{L}(\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_m) = \mathcal{R}\left(\left[\widetilde{A}; \widetilde{b}\right]\right). \tag{29}$$

Furthermore, the real source location $X \in \mathcal{R}(A)$, since $AX = b$. Therefore, it is important to analyze the distance between $\mathcal{R}(\widetilde{A})$, $\mathcal{R}([\widetilde{A}; \widetilde{b}])$ and $\mathcal{R}(A)$. Assumeing that A and $\widetilde{A}$ have full rank, we can obtain the following corollary according to Theorem 2.

Corollary 1. *One has*

$$\begin{aligned} \operatorname{dist}\left(\mathcal{R}(A), \mathcal{R}\left(\widetilde{A}\right)\right) &\leq \frac{\|\Delta A\|}{\sigma_n'}, \\ \operatorname{dist}\left(\mathcal{R}(A), \mathcal{R}\left(\left[\widetilde{A}; \widetilde{b}\right]\right)\right) &\leq \frac{\|\Delta A\|}{\widetilde{\sigma}_n}. \end{aligned} \tag{30}$$

The interlacing theorem (see [31]) implies that

$$\sigma_1' \geq \widetilde{\sigma}_1 \geq \cdots \geq \sigma_n' \geq \widetilde{\sigma}_n. \tag{31}$$

Therefore,

$$\frac{\|\Delta A\|}{\widetilde{\sigma}_n} \leq \frac{\|\Delta A\|}{\sigma_n'}. \tag{32}$$

From Corollary 1 we can see that the upper bound of the distance between $\mathcal{R}(\widetilde{A})$ and $\mathcal{R}(A)$ is smaller than the upper bound of distance between $\mathcal{R}([\widetilde{A}; \widetilde{b}])$ and $\mathcal{R}(A)$. This implies that the TLS solution is expected to be closer to its corresponding unperturbed subspace. Hence, the TLS solution is expected to be more accurate than the LS solution.

5. Simulations

5.1. Simulation Setup. In this section, simulations are carried out to show the effectiveness of the proposed method. Two scenarios are investigated for their effects on localization performance, including Gaussian distribution and truncated Gaussian distribution for measurement noise and sensor node location error. In our simulation, unless otherwise specified, sensors are randomly deployed in a 5000 m × 5000 m area with uniform distribution. An example is shown in Figure 1, where 30 nodes are uniformly distributed. The source is denoted as a triangle with red and the reference node is denoted as a square in blue. The exact location of the reference node can be known.

In order to show the improved performance of our proposed method with the existing closed-form method, we investigate LS method in [16] and WLS method in [17]. Root mean square error (RMSE) is used as the criterion for localization performance. The following simulations are performed with 500 Monte Carlo trials. In all figures, the red solid line with square, black dash line with circle, and blue dotted dash line with triangle represent the results of TLS, WLS, and LS solutions, respectively.

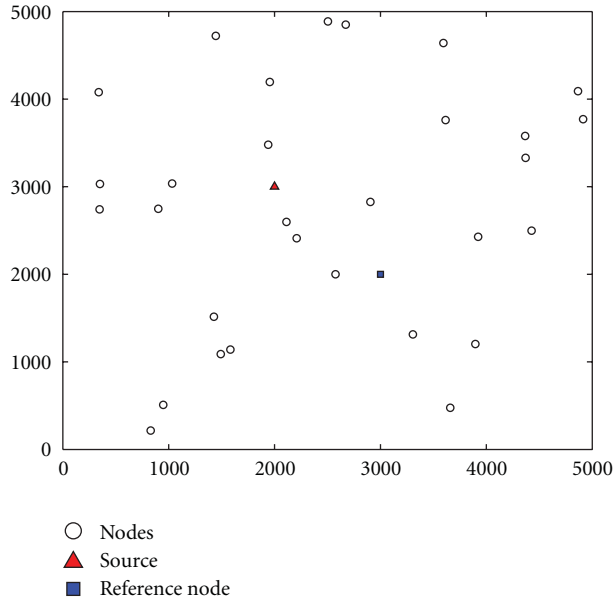


Figure 1: Simulation scenario.

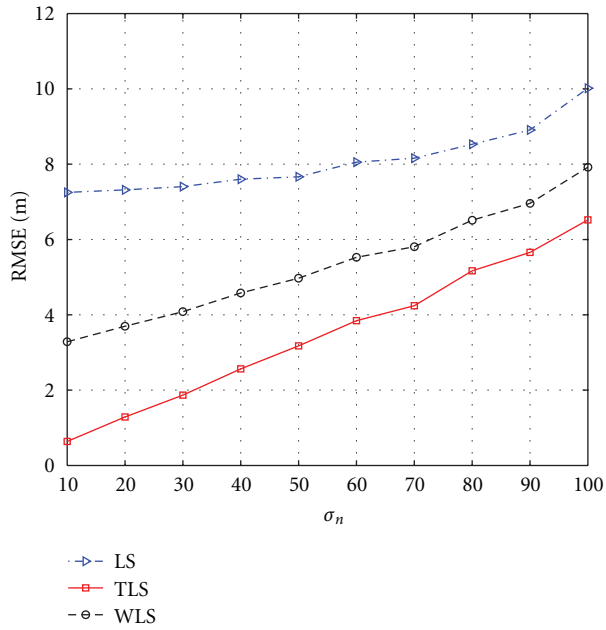


Figure 2: Localization errors versus noise variance (Gaussian case).

5.2. Gaussian Case. In this case, we consider the Gaussian distribution for both measurement noise and sensor node location error. The effect on estimation performance for different measurement noise variance value is investigated. Measurement noise variance (σ_n) is changed in the range 10–100. The variance of perturbation (σ_p) is set to be a constant number 25. The localization performance variation with noise variance is depicted in Figure 2. As noise variance increases, localization performance degrades in all LS, WLS, and TLS algorithms. The TLS algorithm outperforms the LS and WLS algorithms.

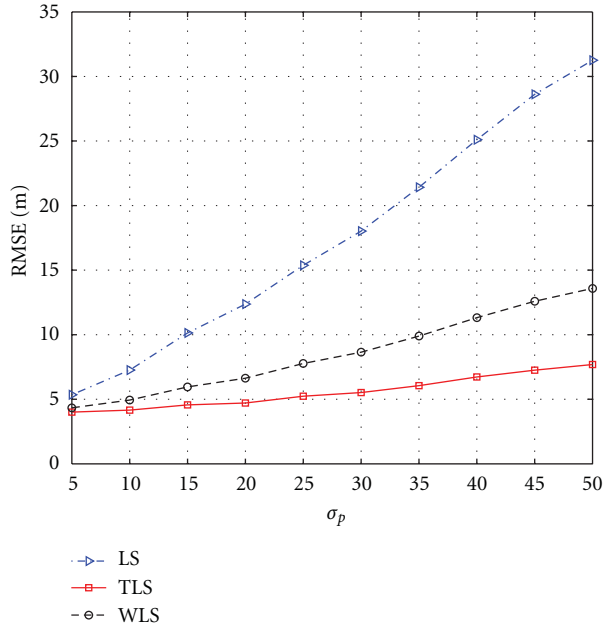


Figure 3: Localization errors versus perturbation variance (Gaussian case).

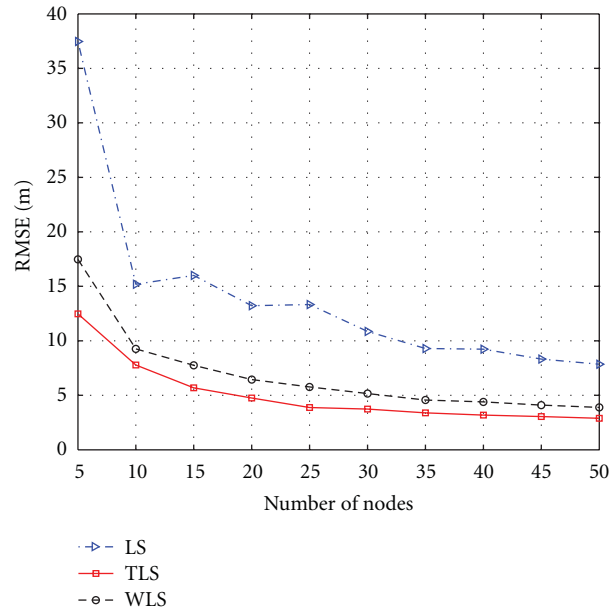


Figure 4: Localization errors versus number of nodes (Gaussian case).

In the second simulation, the effect on estimation performance with different variance of perturbation is investigated. The variance of perturbation is set to be varied from 10 to 100 and sensors are deployed uniformly. The measurement noise variance is set to be 50. The localization performance variation with variance of perturbation is depicted in Figure 3. The TLS algorithm outperforms the LS and WLS significantly with high-level variance of perturbation.

The estimation performance of three algorithms is also investigated with different numbers of deployed nodes. Figure 4 shows the RMSEs versus the number of nodes. Fifty nodes are uniformly generated in the field. Localization starts by using five randomly selected nodes for each algorithm, followed by adding more nodes of five in a group until all fifty nodes are used. The number of FLOPS is considered to evaluate the algorithm computational complexity with variation of the number of nodes. Results are given in Figure 5 with 2-dimensional source location parameter and 5–50 nodes. It can be seen that LS required the fewest FLOPS while TLS required the most FLOPS with SVD of matrix. Although the estimation performance is improved with more nodes for all algorithms, the computation overhead is also increased. t is to show that LS method is the most computationally efficient, whereas WLS method is two-stage least square method and TLS method has singular value decomposition for matrix.

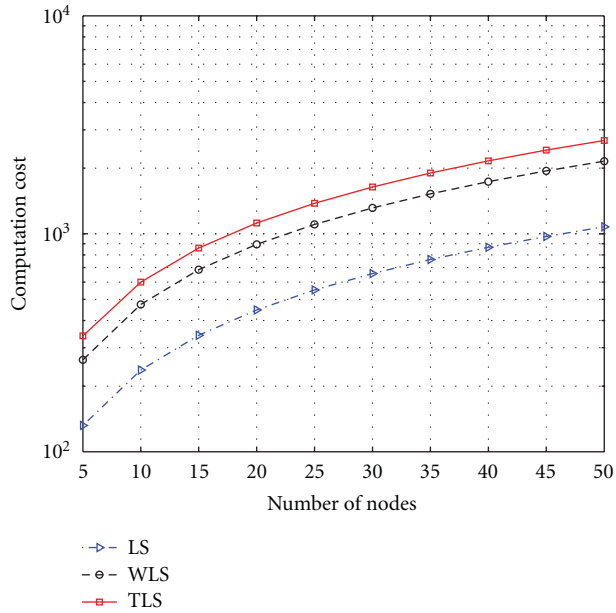


Figure 5: Computation cost versus number of nodes.

5.3. Truncated Gaussian Case. Since the white and Gaussian assumptions are unrealistic in many applications, we consider the truncated Gaussian distribution for both measurement noise and sensor node location error in this case. Let X be a random variable with zero mean Gaussian distribution truncated in the interval $x \leq \alpha\sigma$; its probability density function (pdf) is given by

$$p(x) = \begin{cases} \dfrac{b}{\sqrt{2pi\sigma}}\left(\exp\dfrac{-x^2}{2\sigma^2} - \exp\dfrac{-(\alpha x)^2}{2\sigma^2}\right), & |x| \leq \alpha\sigma, \\ 0, & |x| \geq \alpha\sigma, \end{cases} \tag{33}$$

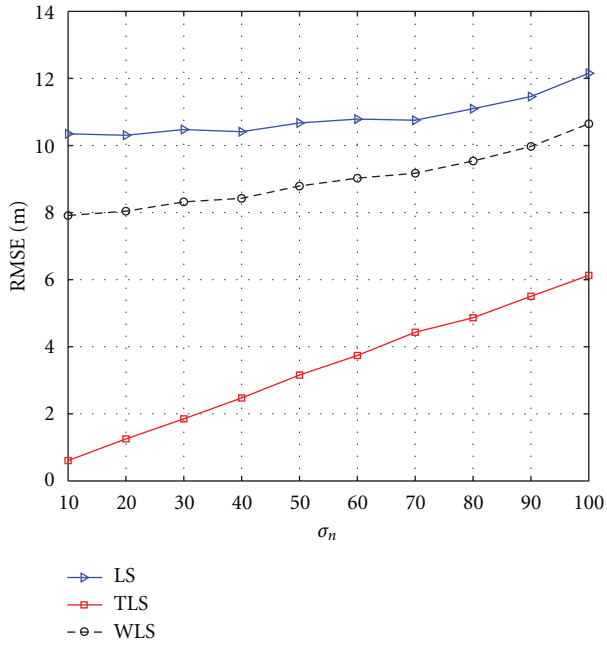


Figure 6: Localization errors versus noise variance (truncated Gaussian case).

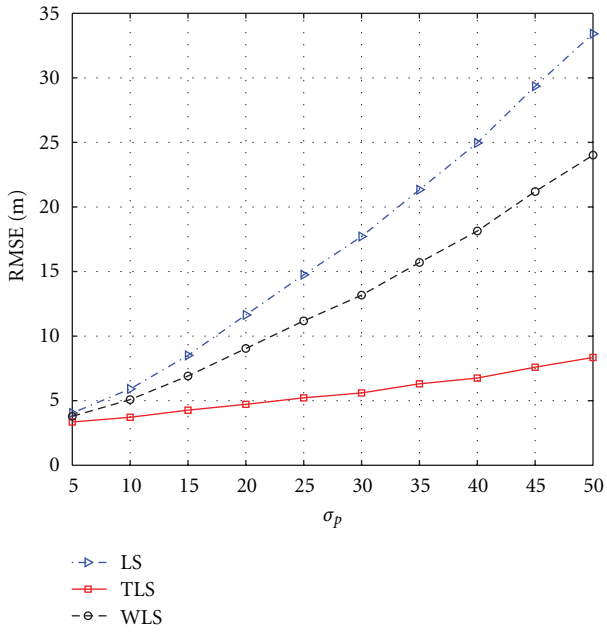


Figure 7: Localization errors versus perturbation variance (truncated Gaussian case).

where b is a normalizing constant, σ^2 is the variance, and α is the factor to extend the interval that the random variable X lies at.

Compared with the Gaussian case, similar results can be obtained and are shown in Figures 6, 7, and 8. The estimation performance for the WLS algorithm distinctly degrades significantly. This is because the weight matrix for this WLS method is calculated according to the Gaussian

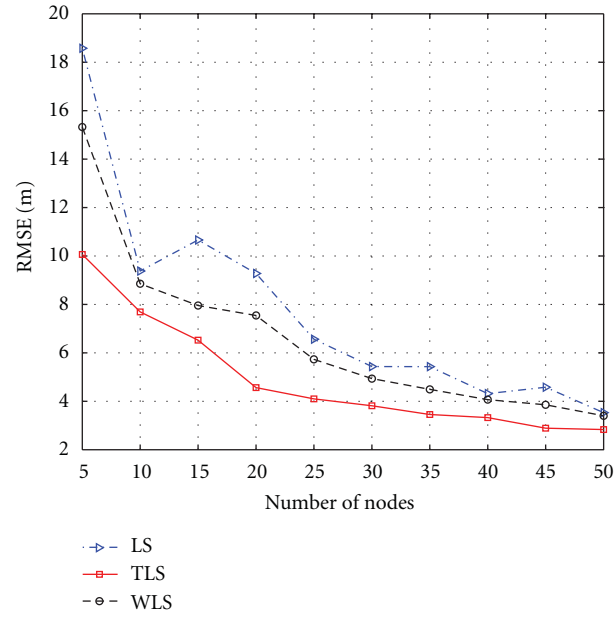


Figure 8: Localization errors versus number of nodes (truncated Gaussian case).

assumption for the measurements noise and sensor node location error. The proposed TLS algorithm still maintains good estimation performance.

6. Conclusions

In this paper, the TDOA model for source localization in sensor networks has been considered. The total least squares (TLSs) algorithm has been developed for location estimation of a stationary source with sensor location uncertainty in which the uncertainty of the sensor location has been formulated as a perturbation. The sensitivity of the TLS solution has also been analyzed to show the advantages of our proposed algorithm. Compared with the existing methods which need the Gaussian assumption for both measurements noise and sensor location error, the TLS approach does not depend on any assumed distribution of the noise and errors. Simulation results show the superior performance of the proposed method.

References

[1] A. Hande, D. Yunus, and E. Cem, "Wireless healthcare monitoring with RFID-enhanced video sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2010, Article ID 473037, p. 10, 2010.

[2] F. Xia, L. Ma, J. Dong, and Y. Sun, "Network qos management in cyber-physical systems," in *Proceedings of the International Conference on Embedded Software and Systems Symposia (ICESS '08)*, pp. 302–307, IEEE, 2008.

[3] Y. Chan, H. Hang, and P. Ching, "Exact and approximate maximum likelihood localization algorithms," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 1, pp. 10–16, 2006.

[4] E. Weinstein, "Optimal source localization and tracking from passive array measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 1, pp. 69–76, 1982.

[5] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[6] J. Chen, K. Yao, and R. Hudson, "Source localization and beamforming," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, 2002.

[7] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[8] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Transactions on Information Theory*, vol. 19, no. 5, pp. 608–614, 1973.

[9] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 5, pp. 1210–1218, 1983.

[10] K. Cheung, H. So, W. Ma, and Y. Chan, "A constrained least squares approach to mobile positioning: algorithms and optimality," *Eurasip Journal on Applied Signal Processing*, vol. 2006, Article ID 20858, p. 23, 2006.

[11] K. Yang, J. An, X. Bu, and G. Sun, "Constrained total least-squares location algorithm using time-difference-of-arrival measurements," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 3, Article ID 5342476, pp. 1558–1562, 2010.

[12] B. Friedlander, "A passive localization algorithm and its accurancy analysis," *IEEE Journal of Oceanic Engineering*, vol. 12, no. 1, pp. 234–245, 1987.

[13] H. C. Schau and A. Z. Robinson, ""Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, pp. 1223–1225, 1987.

[14] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[15] Y. T. Chan and K. C. Ho, "Simple and efficient estimator for hyperbolic location," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 1905–1915, 1994.

[16] M. D. Gillette and H. F. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival," *IEEE Signal Processing Letters*, vol. 15, pp. 1–4, 2008.

[17] K. C. Ho, X. Lu, and L. Kovavisaruch, "Source localization using TDOA and FDOA measurements in the presence of receiver location errors: analysis and solution," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 684–696, 2007.

[18] K. Yang, G. Wang, and Z. Q. Luo, "Efficient convex relaxation methods for robust target localization by a sensor network using time differences of arrivals," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2775–2784, 2009.

[19] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

[20] A. Renaux, P. Forster, E. Boyer, and P. Larzabal, "Unconditional maximum likelihood performance at finite number of samples and high signal-to-noise ratio," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2358–2364, 2007.

[21] K. K. Mada, H. C. Wu, and S. S. Iyengar, "Efficient and robust EM algorithm for multiple wideband source localization," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 6, pp. 3071–3075, 2009.

[22] U. Theodor, U. Shaked, and C. E. de Souza, "A game theory approach to robust discrete-time H^{∞}-estimation," *IEEE Transactions on Signal Processing*, vol. 42, no. 6, pp. 1486–1495, 1994.

[23] S. van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, Society for Industrial Mathematics, 1991.

[24] X. Zhang, *Matrix Analysis and Application*, Tsinghua University Press and Springer, 2004.

[25] I. Markovsky and S. van Huffel, "Overview of total least-squares methods," *Signal Processing*, vol. 87, no. 10, pp. 2283–2302, 2007.

[26] G. Golub and C. van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Md, USA, 1996.

[27] P. Wedin, "Perturbation theory for pseudo-inverses," *BIT Numerical Mathematics*, vol. 13, no. 2, pp. 217–232, 1973.

[28] G. Stewart, "Error and perturbation bounds for subspaces associated with certain eigenvalue problems," *SIAM Review*, vol. 15, no. 4, pp. 727–764, 1973.

[29] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM Journal on Numerical Analysis*, vol. 7, no. 1, pp. 1–46, 1970.

[30] P. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT*, vol. 12, no. 1, pp. 99–111, 1972.

[31] R. C. Thompson, "Principal submatrices IX: interlacing inequalities for singular values of submatrices," *Linear Algebra and Its Applications*, vol. 5, no. 1, pp. 1–12, 1972.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 650387, 10 pages
doi:10.1155/2011/650387

Research Article

Bayesian Network-Based High-Level Context Recognition for Mobile Context Sharing in Cyber-Physical System

Han-Saem Park, Keunhyun Oh, and Sung-Bae Cho

Department of Computer Science, Yonsei University, 262 Seongsanno, Sudaemoon-gu, Seoul 120-749, Republic of Korea

Correspondence should be addressed to Sung-Bae Cho, sbcho@yonsei.ac.kr

Received 10 February 2011; Revised 8 June 2011; Accepted 11 July 2011

With the recent proliferation of smart phones, they become useful tools to implement high-confidence cyber-physical systems. Among many applications, context sharing systems in mobile environment attract attention with the popularization of social media. Mobile context sharing systems can share more information than web-based social network services because they can use a variety of information from mobile sensors. To share high-level contexts such as activity, emotion, and user relationship, a user had to annotate them manually in previous works. This paper proposes a mobile context sharing system that can recognize high-level contexts automatically by using Bayesian networks based on mobile logs. We have developed a ContextViewer application which consists of a phonebook and a map browser to show the feasibility of the system. Experiments of evaluating Bayesian networks and performing the SUS test confirm that the proposed system is useful.

1. Introduction

Recent advancement of ubiquitous sensing, mobile and embedded computing, and wireless communication leads to a new system called cyber-physical systems (CPSs) [1]. They are combination of computation, networking, and physical dynamics where embedded devices are massively networked in order to sense, observe, and control the physical environment. Mobile devices like smart phones can be a good tool for this cyber-physical system since they include various sensors, work as a small computer, and communicate with other devices in wireless networks. Moreover, the number of smart phone users still has been growing dramatically.

On the other hand, a number of web-based social network services including Facebook (http://www.facebook.com/), MySpace (http://www.myspace.com/), and LinkedIn (http://www.linkedin.com/) have been deployed and got explosive popularity with the dissemination of the Internet and social media. Social network sites offer a new way to make and maintain relationships [2]. Because of popular use of smart phones, interest in mobile context sharing systems is increasing for helping social interaction like social network sites. Mobile devices have a variety of sensors like a GPS receiver and a Bluetooth sensor to catch the surrounding and personal contexts [3]. These sensors enable users to communicate with friends by sharing their photos, locations, and so on. Based on this feature, many applications are developed for social network services in mobile environment.

Though a number of mobile context sharing systems are introduced, they cannot generate user's high-level contexts such as activity and emotion automatically. These high-level contexts are more useful for various practical services than easily obtainable low-level information. Previous works should annotate them manually to collect and share this information. This manual annotation is time consuming and can produce incorrect information. Some methods infer high-level contexts based on mobile logs, but they are not adapted to context sharing systems practically. It is required for a system to use a way to infer high-level context considering characteristics of mobile computing in terms of uncertainty, capacity, and power.

This paper presents a system that can recognize users' high-level context based on information collected from mobile devices and develops an application that can share the context based on social relationship among users. The system collects mobile logs, infers high-level contexts with Bayesian networks (BNs), and realizes an application for a smart phone. This system uses a server-client model because a server can cope with the limitation of capacity

and battery power of a smart phone. ContextViewer is a user interface for smart phone using the system, which consists of a phonebook and a map browser to effectively deliver user's context. As sharing users' contextual information more easily and comfortably, the system lets them know their friends' situations and facilitates more communication.

2. Related Works

2.1. Mobile Context Sharing. Sharing contextual information in mobile environment is a hot issue in building social relationships. Many researchers have studied how to construct sensor platform, collect and preprocess features, and design system. ContexPhone is a software platform for context-aware applications. This system acquires context data such as location (Cell ID and GPS) or phone usage from sensors. ContextContacts, one of the applications with this platform, lets users automatically represent and exchange context information [4]. SharedLife is introduced as a generic and reusable content-sharing framework. It collects data from a variety of sensors, stores the information as a set of semantic knowledge models for the user's digital memory, and enables the user to share these memories with others [5]. MyWorld is a context-aware and social networking application. It is focused on activity sharing, real-time location sharing, and media sharing between phonebook contacts [6]. In these previous works, they attempted to recognize and deliver only the low-level contexts. If a user wants to represent their high-level contexts such as activity and emotion, it is required to annotate them manually. The proposed system uses the high-level context recognition model to overcome this problem.

Nomatic is designed for creating content passively by focusing on status messages in instant-messaging (IM) clients as short customizable phrases like "at lunch". Using these cue phrases and sensors, Nomatic recognizes user's activity, place, and other high-level information based on machine learning [7]. While this system supports indoor and laptop environments based on APs (Access Points), it does not cover outdoor and mobile environments sufficiently. Besides, when the system does not have sufficient status message, it is difficult to infer contexts. Santos et al. proposed a method to recognize user activities like walking, running, idle and resting, and presented context sharing applications for instant messenger service, hi5, and microblog service, twitter [8]. Their contexts, however, are very simple and restricted. With advancement of SNS (social network service), context sharing services have been proliferating, but most contextual information is restricted comparing to the high-level contexts in the proposed system. In the previous work, we made recognition models of user activity and emotion for context sharing application [9]. This paper considers additional high-level context of relationships between users and constructs mobile social network using the information. Social relationships between users and mobile social network allow the proposed context sharing system not to depend on manual setting any more.

2.2. Bayesian Network-Based Modeling and Recognition. Bayesian network has emerged as a powerful technique for handling uncertainty in complex domains [10]. It is a model of a joint probability distribution over a set of random variables. The Bayesian network is represented as a directed acyclic graph where nodes correspond to variables and arcs correspond to probabilistic dependencies between connected nodes

$$P\langle B, \theta_B\rangle = P(x_1, x_2, \ldots, x_3) = \prod_{j=1}^{n} P(x_i \mid Pa(x_i)). \quad (1)$$

Equation (1) represents Bayesian network formally. B and θ_B mean Bayesian network structure and probabilistic variables, and $P\langle B, \theta_B\rangle$ means a joint probability distribution of this network. A structure can be represented as $B = (V, E)$, where $V = \{x_1, x_2, \ldots, x_n\}$ is a set of nodes and E is a set of arcs. For each x_i, conditional probability distribution can be represented as $P(x_i \mid Pa(x_i))$, and $Pa(x_i)$ represents a parent set of a variable x_i.

There are two approaches to identify structure and parameters of Bayesian network model. The first approach is learning model from the data. Availability of data depends on the problem domain, and learning is a better choice when we have a lot of data. There are well-known algorithms for learning structures and parameters of Bayesian networks. The K2 algorithm presented by Cooper and Herskovits is the most frequently used algorithm to learn the structure of Bayesian network [11], and maximum likelihood estimation is the most general method to learn parameters of Bayesian network based on statistics [12]. The second one is manual modeling based on domain knowledge, which is crucial in modeling Bayesian network. Experts identify the structure and set parameters based on their knowledge, which is very useful because we cannot obtain reliable data in many real-world problems.

Xu et al. designed a user preference Bayesian network model based on empirical analysis and domain knowledge and applied it to a mobile application [13]. Hong et al. modeled hierarchical Bayesian networks manually for mixed-initiative interaction of human and service robot [14]. A few groups presented a systematic procedure for Bayesian network modeling. Marcot et al. described a guideline for Bayesian network design and suggested it to apply to ecological modeling problem [15]. Laskey and Mahoney proposed a Bayesian network engineering method based on the spiral system lifecycle model of software engineering technique [16]. These works presented a general Bayesian network modeling approach and applied it to real-world problems. On the other hand, a few works used a concept of functional module to design Bayesian networks. Neil et al. presented a procedure to construct a large-scale Bayesian network by using the idiom meaning functional module [17]. Marengoni et al. designed Bayesian network for image understanding and used procedural approach based on functional modules to solve complex problems [18].

There have also been various attempts for recognition using Bayesian networks. Hwang and Cho used a modular Bayesian network model to detect significant events from user's mobile life logs [19]. Krause et al. used BNs to recognize user preferences. To provide smart services, they

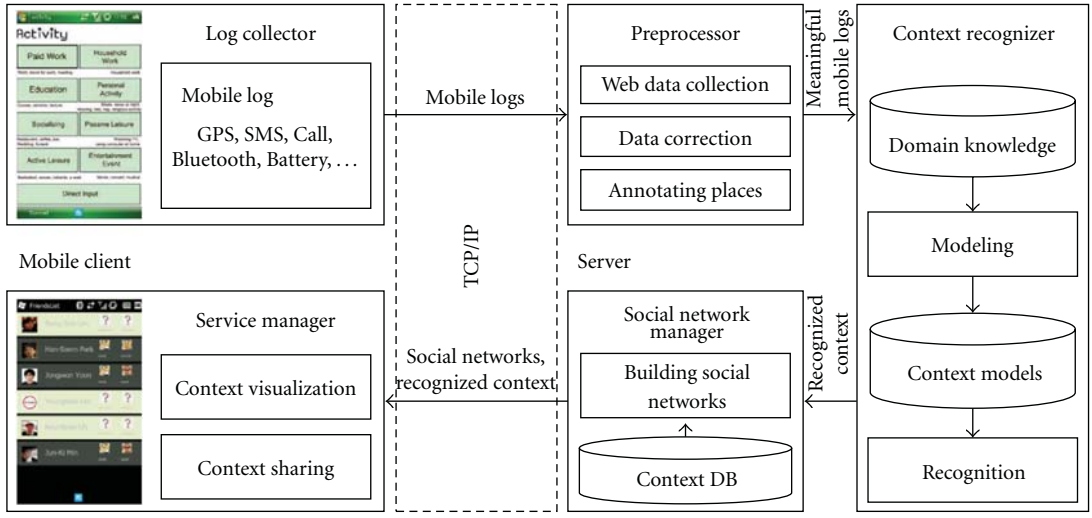


Figure 1: System overview.

clustered log data collected on mobile and wearable sensors, discovered a context classifier that reflected a given user's preferences, and estimated the user's situation [20]. ConaMSN recognized various indoor contexts such as the level of stress, the type of emotion and activity with Bayesian networks and visualized them with a set of icons on a messenger application. The proposed system collected experimental data from smart phones so that we can show the usefulness of the proposed models in uncertain mobile environment [10] and adopted the manual modeling approach for high-level context recognition so that we can overcome the lack of reliable data for learning.

3. Proposed System

We have designed a context sharing system consisting of four modules and user interfaces as shown in Figure 1. Because of the limited resources of smart phone, this system adopts client-server model. When a mobile client connects a server, created is a thread that consists of a preprocessor, a recognizer, and a social network manager. After generating a thread interacted with the client, a log collector collects mobile logs which are low-level contexts from sensors such as GPS coordinates, and call logs. These logs are transported to a server through TCP/IP. A preprocessor generates meaningful logs from web data collected in a server as well as mobile logs collected in mobile clients. A recognizer infers user's high-level contexts of activity, emotion, and relationships between users from refined logs. High-level contexts and meaningful logs are stored in users' context database. A social network manager builds mobile social networks based on high-level contexts including user relationships. A service manager takes a role of presenting contexts to a user and requesting what a user wants to see to a server. When the server receives a request from a mobile client, the server sends contexts to a smart phone. ContextViewer is a user interface to provide a user with information effectively. It consists of two parts: a phonebook and a map browser.

Table 1: Collected mobile life logs.

Source	Type	Attributes
Mobile device	Call	Phone number, type (send/receive/miss), start time, end time
	SMS	Phone number, type (send/receive), message, time
	GPS	Time, latitude, longitude
	Bluetooth	Time, devices nearby
	Battery	Time, be charging, power
Web	Weather	Start time, end time, weather
	Temperature	Start time, end time, temperature

3.1. Log Collector. The log collector module continuously gathers available low-level information such as call logs, SMS logs, GPS (Global Positioning System) coordinates, device status, Bluetooth, and weather from web in a mobile device. Most of context information is sent in the predefined expiration time for each log, but web data are collected in server part to reduce the communication bandwidth and power of mobile device. The data are used to recognize the user's high-level contexts such as what the user is doing and how the user feels.

Call and SMS logs are collected when an event occurs. For example, whenever the user makes call, the smart phone records logs including phone number, call type (send/receive/miss), start time, and end time. The module sends GPS logs and Bluetooth logs to a server periodically. The GPS logs present the places where the user visited. Bluetooth sensor can discover devices which allow the mobile device to collect information on other Bluetooth devices carried by people nearby. Weather log is obtained from web based on location and time. If weather or temperature is changed, the module writes a new log. Table 1 shows the types of logs. The logs for several sources are sorted according to the time.

Table 2: Examples of matching a GPS coordinate to a place in Yonsei University.

Place	Longitude	Latitude
Post office in Seodeamoon-gu	126.93204267325	37.560804484469
Engineering building	126.936001667	37.560815
Subway station in Sinchon-dong	126.9368577113	37.555257212296
Library	126.937158333	37.563678333

3.2. Preprocessor. The preprocessor module is in charge of producing meaningful information from raw mobile logs. Preprocessing step includes correcting data and annotating places. Data correction is to correct wrong data because of no signal. When GPS signals are not available, it writes a GPS log which has a value of the latest coordinate received from satellites. Because one of the purposes of the proposed system is to share information in real time, this approach is not used, though it is possible to infer locations by using previous coordinates and next coordinates. Place is labeled by the module. GPS logs, Bluetooth logs, and time are important traits for this labeling. If a GPS device in a smart phone detects coordinates, a user locates in outdoor. In outdoor environment, the name of a place can be labeled by matching near GPS latitude and longitude with a location in the predefined location table. Examples of matching a GPS coordinate to a place in Yonsei University are shown in Table 2. Bluetooth logs can be important clues to infer a location in indoor environment. For example, if a Bluetooth receiver catches Bluetooth ids of coworkers' computers, the user is in his workplace. Locations annotated are used to infer high-level contexts.

3.3. Context Recognizer. The context recognizer models high-level contextual information to recognize. Here, high-level contexts include user activity, emotion, and relationship between users. Table 3 shows the available values for each context. As data from sensors on a smart phone are often missing, uncertain, and incomplete, it is required to deal with uncertainty and missing values. BNs can address such problems by providing a robust inference based on probability. In this paper, the recognizer uses BN to model and recognize user's high-level contexts. We have designed Bayesian network models based on domain knowledge for general users. Modeling process includes designing the structure and parameters. For inference, we have used a well-known BN library SMILE (http://genie.sis.pitt.edu/).

In order to recognize activity, BNs consist of five factors: mobile device status, spatial, temporal, environmental, and social factors. These are shown in Table 4 together with the corresponding variables. The recognizer calculates the probability of each activity at present based on these modules. Each activity is differently influenced on each module. For example, a temporal factor module has an effect on regular actions such as "meal" and "sleep," but an environmental factor module does not. "Sport" is sensitive to weather that is an environmental factor module. Figure 2 shows an example of BN which is a model to infer "Sport" activity.

Table 3: The available values for high-level contexts used.

Type	Value
Activity	Move, study, meal, sleep, sport, play, rest
Emotion	Bored, contented, excited, happy, nervous, normal, relaxed, sad, upset
User relationship	For private relationship—close friend, friend, acquaintance, none for work relationship—close colleague, colleague, acquaintance, none

Table 4: Five factors and the corresponding variables of activity recognition BNs.

Factor	Variables
Mobile device status factor	Call and SMS, power
Spatial factor	Place, detailed place
Temporal factor	Holiday, time zone
Environmental factor	Temperature, weather
Social factor	Occupation, sex

Table 5: Arousal-valence value and the corresponding emotion.

Emotion	Arousal	Valence
Excited	0.8	0.6
Happy	0.6	0.8
Contented	0.4	0.8
Relaxed	0.2	0.6
Bored	0.2	0.4
Sad	0.4	0.2
Upset	0.6	0.2
Nervous	0.8	0.4

BNs to infer emotion use the result of activity inference BNs since activity has an influence on user's emotion directly. Inferred results of BNs are represented as axes of arousal and valence. It is based on arousal-valence emotion value defined as Table 5. Figure 3 shows BNs designed to recognize emotion. The probability of arousal and valence indicates the nearest feeling in Table 5. Figure 4 illustrates the BN to infer the relationship between users. It infers private and work relationships between two users.

Figure 4 provides a Bayesian network to infer the relationship between two users. It uses the activity and emotion inferred as well as other mobile logs of call, SMS, common schedule, and proximity based on Bluetooth information and user activities. If two users are searched by the device of the other user and their activities are private, it is set as P_Proximate. If activities are work related in the same condition, it is set as W_Proximate. The variables used for this BN and their possible states are summarized in Table 6. This Bayesian network is modeled to infer two types of relationships: private and work relationships. Inferred result is used to make mobile social networks by social network manager in the next step.

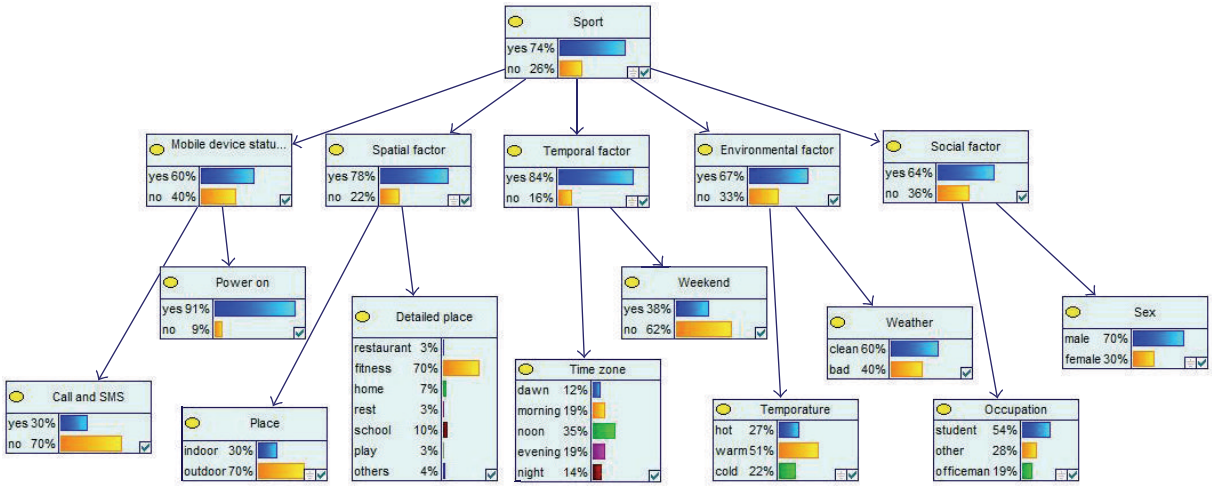


Figure 2: A Bayesian network to infer "Sport" in the recognizer.

Table 6: Variables and the possible states of Bayesian network model designed.

Variable	State
Private relation	Close friend, friend, acquaintance, none
Work relation	Close colleague, colleague, acquaintance, none
Contact related	High, low
Emotion related	High, low
Social activity related	High, low
Work activity related	High, low
Call frequency	Many, some, none
SMS frequency	Many, some, none
Call duration	Long, short
Emotion	Positive, negative
Common social activity	Many, some, none
Proximity	PProximate, WProximate, NotProximate
Common work activity	Many, some, none
Common schedule	Yes, no

3.4. Social Network Manager. A social network manager constructs a mobile social network using mobile logs and inferred high-level contexts. Generally, a relationship between users in a mobile social network considers only the existence (or strength) of connection. This paper identifies semantic relationships between users and builds the mobile social network with them. These relationships are inferred from the Bayesian network in Figure 4.

The mobile social network $\mathcal{N}$ is defined as $\mathcal{N} = \langle \mathcal{U} \cdot \mathcal{R} \rangle$, where $\mathcal{U}$ is a set of users $\{u_1, u_2, \ldots, u_n\}$, and $\mathcal{R}$ is a set of relationships between users. Here, n is the number of users. Our model infers two kinds of relations, and $\mathcal{R}$ is defined according to the type of relationship as follows:

$$\begin{aligned} &\mathcal{R}_{\text{private\_relation}} \\ &\quad = \{\text{close friend, friend, acquaintance, none}\}, \\ &\mathcal{R}_{\text{work\_relation}} \\ &\quad = \{\text{close colleague, colleague, acquaintance, none}\}. \end{aligned} \tag{2}$$

Elements in these relationships are relation candidates inferred from Bayesian network model.

We also need to decide how mobile social networks can be displayed. This network is represented as a graph where the nodes correspond to mobile users and the links correspond to relationships between them. The type of relationship mentioned above decides its thickness, and the importance of users decides their size in a graph. To calculate this importance, the closeness centrality is used as follows [21]:

$$C(u_i) = \frac{n-1}{\sum_{j=1}^{n} d\left(u_i, u_j\right)}. \tag{3}$$

A function $d(u_i, u_j)$ measures a geodesic distance between two users, and it is modified as follows:

$$d\left(u_i, u_j\right) = \begin{cases} \text{if } u_i \text{ and } u_j \text{ are adjacent and} & \\ \quad \text{the relation is } \textit{closefriend} & \\ \quad \text{or } \textit{close colleague}: & 0.25 \\ \quad \text{the relation is } \textit{friends} \text{ or } \textit{colleague}: & 0.5 \\ \quad \text{the relation is } \textit{acquaintance}: & 0.75 \\ \quad \text{the relation is } \textit{none}: & 1 \\ \text{if } u_i \text{ and } u_j \text{ are not adjacent:} & \\ \sum_{k=1}^{\text{number of relations between } u_i \text{ and } u_j} d(u_{i+k-1}, u_{i+k}). & \end{cases} \tag{4}$$

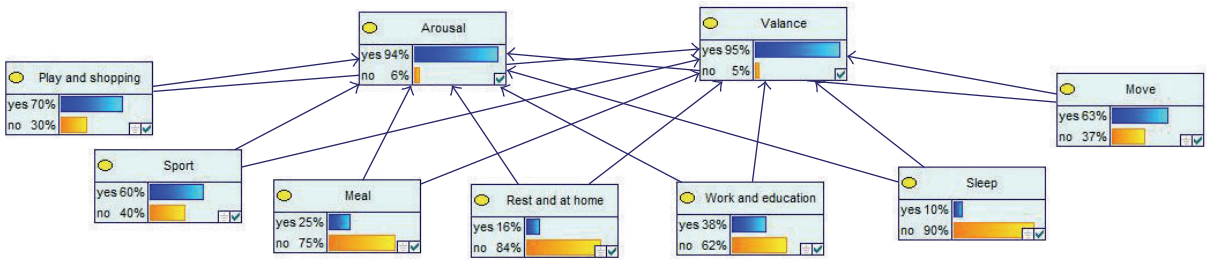


Figure 3: A Bayesian network to infer emotion (Inferred result corresponds to "Excited").

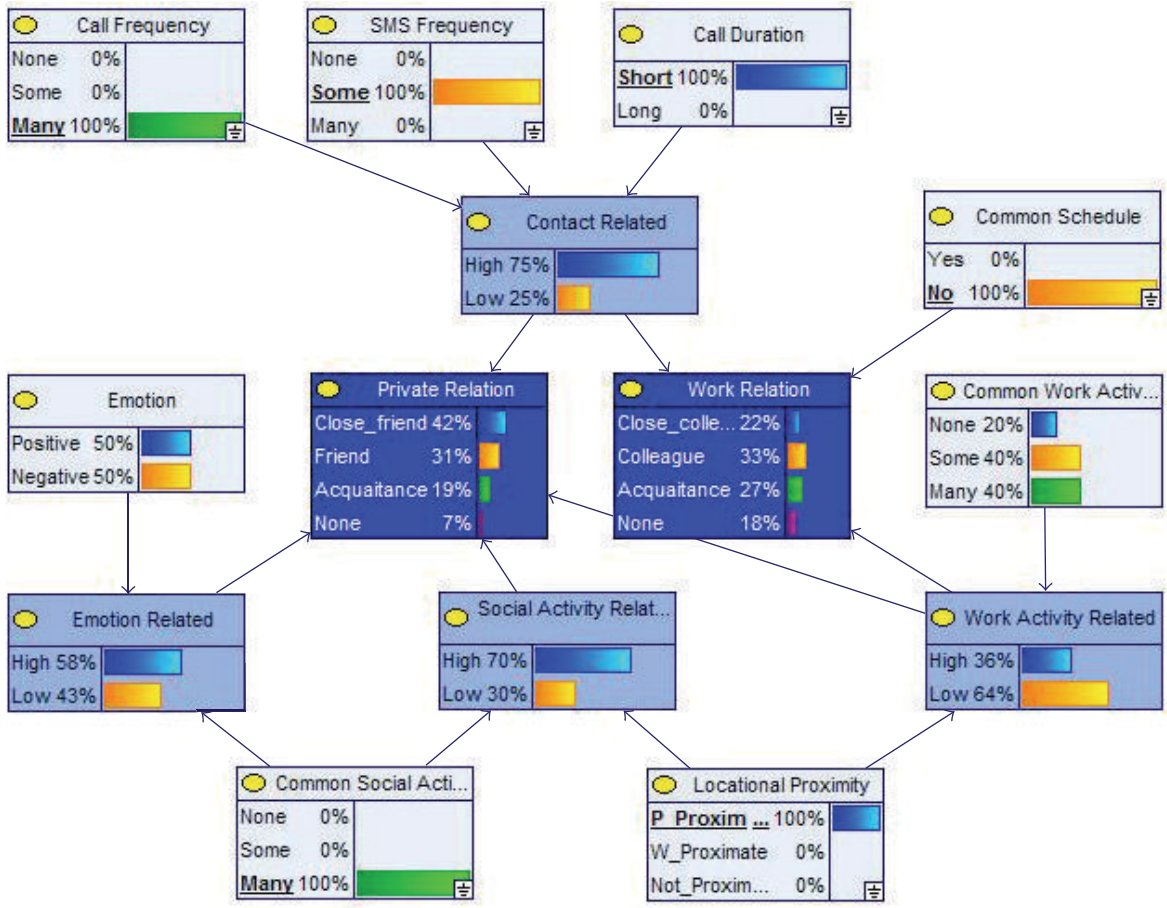


Figure 4: A Bayesian network to infer the relationship between users.

We also restrict the maximum number of links between two users as three when calculating the distance function. It means that we would neglect the closeness if the distance of two users is very far. It is realistic because the influence of a certain user to another degrades as the number of users in between grows.

3.5. Service Manager for Context Sharing. In a mobile client, a service manager provides the context sharing service to users based on the mobile social network made by the social network manager. Context sharing service includes sharing information of friends registered in user's phonebook. Also, it gets map images of locations that GPS coordinates indicate from Google Maps (http://maps.google.com/). A contact list in a phonebook can be a mobile social network centering users themselves potentially. The proposed service provides the contact list with user name, image, activity, and emotion if users in that list are close enough. That is, private information like activity or emotion is shared between close friends. An application of this context sharing service will be discussed in the experiments later.

4. Experiments

4.1. Environment and Data. The proposed system has a server written in C# (.NetFramework3.5) based on Windows

Table 7: A part of phone-call log of User 1.

User ID	Start time	End time	Status	Receiver ID
010-4135-xxxx	2009-09-14 20:59	2009-09-14 20:59	Sended	010-7190-xxxx
010-4135-xxxx	2009-09-14 20:54	2009-09-14 20:54	Sended	010-8563-xxxx
010-4135-xxxx	2009-09-15 11:26	2009-09-15 11:26	Sended	010-7190-xxxx
010-4135-xxxx	2009-09-16 18:18	2009-09-16 18:18	Received	010-2085-xxxx
010-4135-xxxx	2009-09-16 18:08	2009-09-16 18:09	Received	010-7182-xxxx
010-4135-xxxx	2009-09-16 16:10	2009-09-16 16:10	Missed	010-5378-xxxx
010-4135-xxxx	2009-09-16 16:09	2009-09-16 16:09	Sended	010-5378-xxxx
010-4135-xxxx	2009-09-16 18:18	2009-09-16 18:18	Received	010-2085-xxxx
010-4135-xxxx	2009-09-16 18:08	2009-09-16 18:09	Received	010-7182-xxxx
010-4135-xxxx	2009-09-16 16:10	2009-09-16 16:10	Missed	010-5378-xxxx
010-4135-xxxx	2009-09-16 16:09	2009-09-16 16:09	Sended	010-5378-xxxx
010-4135-xxxx	2009-09-17 15:28	2009-09-17 15:31	Sended	031-299-xxxx

Table 8: A part of an activity log of User 1.

User ID	Start time	End time	Activity
010-4135-xxxx	2009-09-15 10:55	2009-09-15 12:18	Study (work)
010-4135-xxxx	2009-09-15 12:19	2009-09-15 12:32	Meal
010-4135-xxxx	2009-09-15 12:33	2009-09-15 13:12	Unknown
010-4135-xxxx	2009-09-15 13:33	2009-09-15 13:32	Rest
010-4135-xxxx	2009-09-15 15:56	2009-09-15 15:55	Study (work)
010-4135-xxxx	2009-09-15 15:35	2009-09-15 17:37	Move
010-4135-xxxx	2009-09-15 17:38	2009-09-15 18:31	Study (work)
010-4135-xxxx	2009-09-15 18:32	2009-09-15 20:03	Meal
010-4135-xxxx	2009-09-15 20:04	2009-09-15 21:24	Unknown
010-4135-xxxx	2009-09-15 21:25	2009-09-15 23:09	Move
010-4135-xxxx	2009-09-15 23:10	2009-09-15 23:22	Rest
010-4135-xxxx	2009-09-15 23:23	2009-09-15 23:59	Sleep

platform and mobile clients that contain a mobile log collector and ContextViewer written in C# (.NetCompactFramework3.5) using SAMSUNG T*Omnia SCH-M490, a smart phone based on Windows mobile 6.5. MySQL 5.1 is used as a database management system. The server and the mobile clients communicate with each other through TCP/IP socket.

In order to model and evaluate Bayesian networks appropriately, we collected mobile logs from eleven graduate students for three weeks. Participants consist of one female student and ten male students, and eight students have experiences of using smart phones before while the others have not. We asked them to annotate high-level context information of activity and emotion manually. Three weeks seem to be too short, but Eagle et al. analyzed that long-term data from mobile devices have enormous redundancy and observation for two weeks can largely replicate the data produced from nine months observations [22].

Tables 7 and 8 show parts of mobile logs collected by a user (User 1). Table 7 provides a phone call log, which includes sender, start time, end time, status, and receiver. The last four digits of cell phone numbers are replaced by "xxxx" for the sake of privacy. Table 8 shows an activity log of the same user. Activity and emotion have been labeled together with start and end time.

4.2. Evaluation of Bayesian Networks. In order to evaluate Bayesian networks, inference models in the system, we calculate the recall and precision of Bayesian networks for activity inference by using collected mobile logs of 11 individuals. Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. The result is shown in Figure 5. The probability is 80∼95% about "study," "meal," and "sleep." However, "play" and "rest" are low because these activities are done irregularly and regardless of environment. It is also unusual that an activity "meal" has a low recall rate while it has a high precision. It is because "meal" happens both indoor and outdoor. Outdoor "meal" can be recognized easily, but indoor "meal" is confused with other indoor activities like rest and study.

The inferred relationships are compared with actual relationships from self-reported data. Users were asked to select one of the predefined relationships to other users. Options are "close friend/friend/acquaintance/none" for private relationship and "close colleague/colleague/acquaintance/none" for work relationship. Answers of two users are not always the same, and in those cases, we decided that the inference was correct if the inferred answer was the same as either of two different answers. Table 9 provides accuracy, precision, and recall of this comparison.

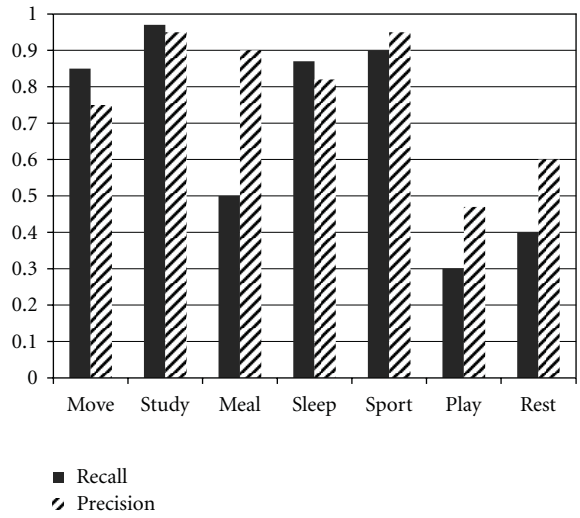


Figure 5: A result of evaluating Bayesian networks for activities.

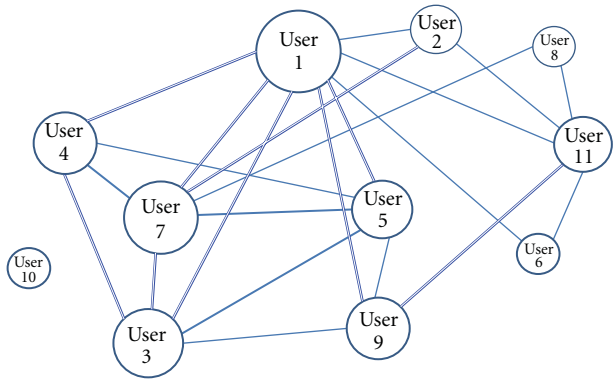


Figure 6: A mobile social network built based on private relations between users.

Table 9: Accuracy in comparison of actual and inferred relationships.

Used model	Accuracy	Precision	Recall
Private relationship-based social network	72.2%	69.6%	69.9%
Work relationship-based social network	64.8%	61.9%	60.9%

4.3. Mobile Context Sharing Application. Using the proximity and inferred high-level contexts together, we can make various mobile social networks. For example, two users are close in private relation if they often watch movie together. Here, we present two networks as examples.

(1) A network based on private relationships: First, we built a mobile social network based on private relationships between users, as shown in Figure 6. Socializing activities like meal and play or active leisure activities like sport took important roles in this network. In this network, the size of a node represents the importance of a corresponding user, and the position is based on their physical location. In reality users 3, 4, 7, and 10 share an office, and the other users share the other office. Thicknesses of links are decided by the relations inferred.

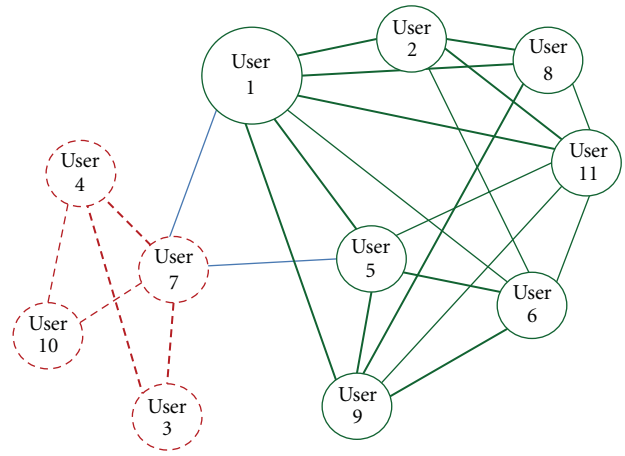
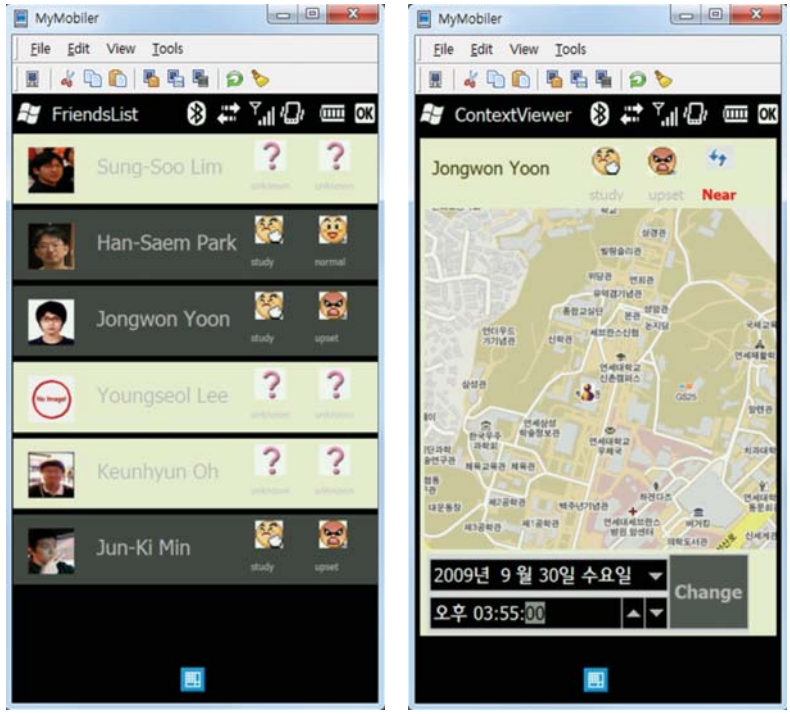


Figure 7: A mobile social network built based on work relations among users.

(2) A network based on work relationships: Figure 7 shows a mobile social network built based on work relationships. Obviously, activities related to work (study) are significant in this network. We found two large groups in the network (solid lines and dotted lines), and it is found that users in the same group work at the same office. In short, this network depends on the proximity very much. The network is visualized with the same method as private relationship-based network, and user 1 is shown as the most important person in this network. In the real world, user 1 is the manager of the lab and a main room. In terms of the relationships, this network depends on the proximity very much.

Figure 8 shows an example of using the system. When the user executes ContextViewer, an application in the mobile client, a phonebook is displayed (Figure 8(a)). In the interface, the user sees only friends who allow the user to view their contexts and the user does for them. The user catches that "Yoon" feels upset. The user wants to talk with him and encourage him. When the user selects his item, a map browser is opened (Figure 8(b)). The user knows where he is and if he is near the user. The user can call him and meet him.

4.4. The Usability Test. In order to validate the usefulness of the proposed system, we performed a subjective test about the implemented application for thirteen subjectives based on the System Usability Scale (SUS) questionnaires. The SUS is a simple, ten-item scale giving a global view of subjective assessments of usability. Table 10 shows ten items for questionnaire, and each subjective should answer the item as 5-degree scale where one means strongly disagree whereas five means strongly agree. Total score from ten answers has a range of zero to one hundred for each subject [23].



(a) Phonebook (b) Map browser

Figure 8: ContextViewer.

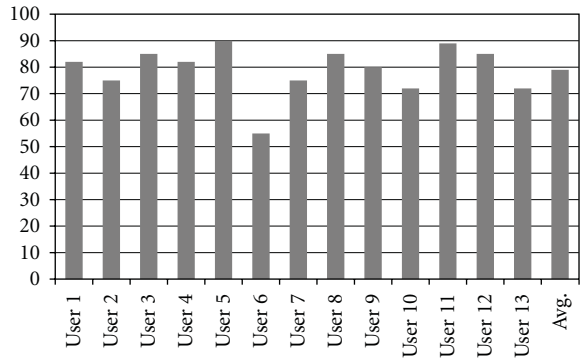


Figure 9: SUS scores for the proposed system.

It can be thought that the score of 50 is neutral, and the system with more than 50 has a good usability. User 6, who gave the lowest SUS score, has proved that he has never used a smartphone. Figure 9 shows the SUS test results of thirteen participants, which indicate that the system provides an effective way for sharing contexts conveniently. By questionnaire, numbers 3 and 5 have been given the best score while number 9 has been given the worst score. Questionnaires with the best score include very important parts (no. 3 easy to use and no. 5 well integrated functions) in evaluating mobile services.

5. Concluding Remarks

In this paper, we have presented a context sharing system in mobile environment that recognizes and shares high-level context information automatically to support high-confidence cyber-physical systems. In the proposed system, a mobile client collected mobile logs, sent them to a server, and provided context sharing interface. ContexViewer is a service to share contextual information of friends. It consists of a phonebook and a map browser. The server preprocesses and recognizes high-level contexts with Bayesian network models to handle uncertainty in mobile environment. When the mobile client requests some contexts, if not permitted, the server prevents to send them. The usefulness of the proposed system is shown by evaluating Bayesian network models that

Table 10: The available values for activity and emotion.

No.	Questionnaire
(1)	I think I would like to use this system frequently
(2)	I found the application unnecessarily complex
(3)	I thought the system was easy to use
(4)	I think that I would need the support of a technical person to be able to use this application
(5)	I found the various functions in this application well integrated
(6)	I thought there was too much inconsistency in this system
(7)	I would imagine that most people would learn to use this system very quickly
(8)	I found the system very cumbersome to use
(9)	I felt very confident using the system
(10)	I needed to learn a lot of things before I could get going with this system

infer activity and user relationship and by performing SUS test of our context sharing application. The proposed system helps users to share context information without manual setting.

There are several ongoing tasks in this work. Bayesian network models designed in this paper are for general users, and it cannot work well for all substantial users. A few more models for different types of users will be helpful to make this model more general. It is also required to compare the proposed context sharing system with the alternatives. Thirdly, we are planning to apply the proposed method to other applications. The ContextViewer only shows the context information, but it can provide more powerful services based on high-level contexts. For example, behavior recommendation for mobile user can be possible based on diverse contexts. Finally, we are using the context information collected from smart phones, but the source of context information can be extended to a smart space, which has diverse sensors in an environment for intelligent services.

Acknowledgment

This research was supported by the Original Technology Research Program for Brain Science through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0018948).

References

[1] W. Wolf, "The good news and the bad news," *Computer*, vol. 40, no. 11, pp. 104–105, 2007.

[2] C. Steinfield, N. B. Ellison, and C. Lampe, "Social capital, self-esteem, and use of online social network sites: a longitudinal analysis," *Journal of Applied Developmental Psychology*, vol. 29, no. 6, pp. 434–445, 2008.

[3] M. Sama, D. S. Rosenblum, Z. Wang, and S. Elbaum, "Multi-layer faults in the architectures of mobile, context-aware adaptive applications," *Journal of Systems and Software*, vol. 83, no. 6, pp. 906–914, 2010.

[4] M. Raento, A. Oulasvirta, P. Renaud, and H. Toivonen, "ContextPhone: a prototyping platform for context-aware mobile applications," *IEEE Pervasive Computing*, vol. 4, no. 2, pp. 51–59, 2005.

[5] A. Kröner, M. Schneider, and J. Mori, "A framework for ubiquitous content sharing," *IEEE Pervasive Computing*, vol. 8, no. 4, Article ID 5280685, pp. 58–65, 2009.

[6] K. Sorathia and A. Joshi, "My world—social networking through mobile computing and context aware application," in *Proceedings of the Communications in Computer and Information Science (CCIS '09)*, vol. 53, pp. 179–188, 2009.

[7] D. J. Patterson, X. Ding, S. J. Kaufman, K. Liu, and A. Zaldivar, "An ecosystem for learning and using sensor-driven im status messages," *IEEE Pervasive Computing*, vol. 8, no. 4, Article ID 5280683, pp. 42–49, 2009.

[8] A. C. Santos, J. M. P. Cardoso, D. R. Ferreira, P. C. Diniz, and P. Chaínho, "Providing user context for mobile and social networking applications," *Pervasive and Mobile Computing*, vol. 6, no. 3, pp. 324–341, 2010.

[9] K. Oh, H. S. Park, and S. B. Cho, "A mobile context sharing system using activity and emotion recognition with Bayesian networks," in *Proceedings of the Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC '10)*, pp. 244–249, 2010.

[10] J. H. Hong, S. I. Yang, and S. B. Cho, "ConaMSN: a context-aware messenger using dynamic Bayesian networks with wearable sensors," *Expert Systems with Applications*, vol. 37, no. 6, pp. 4680–4686, 2010.

[11] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[12] E. B. Anderson, "Asymptotic properties of conditional maximum likelihood estimators," *Journal of Royal Statistics Society Series B*, vol. 32, no. 2, pp. 283–301, 1970.

[13] D. J. Xu, S. S. Liao, and Q. Li, "Combining empirical experimentation and modeling techniques: a design research approach for personalized mobile advertising applications," *Decision Support Systems*, vol. 44, no. 3, pp. 710–724, 2008.

[14] J. H. Hong, Y. S. Song, and S. B. Cho, "Mixed-initiative human-robot interaction using hierarchical Bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 37, no. 6, pp. 1158–1164, 2007.

[15] B. G. Marcot, J. D. Steventon, G. D. Sutherland, and R. K. McCann, "Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation," *Canadian Journal of Forest Research*, vol. 36, no. 12, pp. 3063–3074, 2006.

[16] K. B. Laskey and S. M. Mahoney, "Network engineering for agile belief network models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 4, pp. 487–498, 2000.

[17] M. Neil, N. Fenton, and L. Nielsen, "Building large-scale Bayesian networks," *Knowledge Engineering Review*, vol. 15, no. 3, pp. 257–284, 2000.

[18] M. Marengoni, A. Hanson, S. Zilberstein, and E. Riseman, "Decision making and uncertainty management in a 3D reconstruction system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 852–858, 2003.

[19] K. S. Hwang and S. B. Cho, "Landmark detection from mobile life log using a modular Bayesian network model," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12065–12076, 2009.

[20] A. Krause, A. Smailagic, and D. P. Siewiorek, "Context-aware mobile computing: learning context-dependent personal preferences from a wearable sensor array," *IEEE Transactions on Mobile Computing*, vol. 5, no. 2, pp. 113–127, 2006.

[21] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[22] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 36, pp. 15274–15278, 2009.

[23] J. Brooke, "SUS: a quick and dirty usability scale," in *Usability Evaluation in Industry*, P. W. Jordan et al., Ed., Taylor & Francis, London, UK, 1996.

Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2011, Article ID 717208, 9 pages
doi:10.1155/2011/717208

Research Article

Distributed Algorithm for Traffic Data Collection and Data Quality Analysis Based on Wireless Sensor Networks

Nan Ding, Guozhen Tan, Wei Zhang, and Hongwei Ge

Department of Computer Science and Technology, Dalian University of Technology, Dalian Liaoning 116023, China

Correspondence should be addressed to Nan Ding, dingnan@dlut.edu.cn

Received 14 February 2011; Accepted 10 July 2011

The growing need of the real-time traffic data has spurred the deployment of large-scale dedicated monitoring infrastructure systems, which mainly consist of the use of inductive loop detectors. However, the loop sensor data is prone to be noised or even missed under harsh environment. The state-of-the-art wireless sensor networks provide an appealing and low-cost alternative to inductive loops for traffic surveillance. Focusing on the urban traffic data collection, this paper proposes a distributed algorithm to collect the traffic data based on sensor networks and improve the reliability of data by quality analysis. Considering the certain correlated characteristics, this algorithm firstly processes the data samples with an aggregation model based on the mean filter, and then, the data quality is analyzed, and partial bad data are repaired by the cusp catastrophe theory. The performance of this algorithm is analyzed with a number of simulations based on data set obtain in urban roadway, and the comparative results show that this algorithm could obtain the better performance.

1. Introduction

Nowadays, the traffic jams is a difficult problem that confronts the urban residents with environmental pollution, traffic incident, and great financial loss every year. The intelligent traffic system (ITS) has proved to be the most effective approach to resolve this problem. In the ITS, the real-time traffic data collection and data quality analysis play the critical roles for studying and monitoring the traffic state.

As the ITS operations expand in major urban areas, vast amount of traffic detectors are deployed in road networks which provide the most abundant source of traffic data. But the detectors are prone to errors and malfunctioning, data samples are often missing or invalid. As reported, 30% out of 25000 detectors do not work properly daily in California [1]. And a program named *Performance Measurement System* (PeMS) has been launched for the traffic data health checking by the California Department of Transportation and researchers of the University of California at Berkeley [2, 3]. The PeMS currently collects more than 1GB per day from thousands of loop detectors in the California [4]. According to statistics, only 32.78% of detectors worked well and provided *good data* (reliable data), and the others worked with failure and provided *missing data* or *bad data*. While, some good data reported by the well-functioned detector working in the severe environment such as incident or bad weather were discarded as error data, because they were apparent outliers in normal condition. In fact, these data are more valuable to the ITS in some extent. To improve the usability of traffic data, it is urgent to develop new technology to collect the data and the effective algorithm to identify the real bad data or repair them.

Wireless Sensor Network (WSN) is a distributed collection of sensor nodes having potential application in traffic surveillance system with detection accuracy as good as that of inductive loop detectors. As they have a much higher configuration flexibility, which makes the system scalable and deployable everywhere in the road network, the sensor networks offer an attractive alternative to inductive loops for traffic surveillance [5].

In this paper, we focus on studying the performance of the traffic data detection based on WSN. And a *distributed traffic data detection and quality analysis* (DTDDQA) algorithm is proposed to process the traffic data collection and data quality analysis. This algorithm firstly aggregates the samples according to their correlated characters. Then, based on the nonlinear and catastrophic characters of traffic data, the algorithm analyzes the data quality and repairs partial

bad data based on the cusp catastrophe theory. Furthermore, a number of simulations are conducted to evaluate the performance of the algorithm. The simulation results show that this algorithm outperforms other data quality analysis algorithms with better performance and robustness.

2. Related Work

2.1. Traffic Data Collected by WSN. WSN is an effected technology and a revolution in remote information sensing and collection applications. Sensor node has advantages such as low costs, small size, wireless communication, and high sensing accuracy, and it can be deployed with great quantity. Compared with the traditional centralized data processing mode, WSN provides a new solution for distributed method to acquire and process traffic data [6]. In the prior research publications [5, 7, 8], the possibility of replacing traditional methods with WSN is creatively researched in the *California Partners for Advanced Transit and Highways* (PATH) project of University of California at Berkeley. In this three-year research project, they focused on the prototype design, analysis, and performance of WSN for traffic surveillance using both acoustic and magnetic sensors. And in California, they verified the feasibility of collecting information based on WSN with a large number of on-road experiments. Jeahoon deployed WSN in the road networks and introduced an *autonomous passive localization* (APL) scheme to perform the localization using vehicle-detection timestamps along with the road map of target area. It was evaluated in Minnesota roadways, and the results shown that it was effective [9].

In the state-of-the-art vehicle detection based on WSN, magnetic sensor, such as Honeywell HMC1002, has been integrated with sensor node of WSN. The magnetic sensor can measure the change of Earth's magnetic field with high accuracy [10]. According to the magnetic field distortion signal caused by moving vehicle, an efficient vehicle detection algorithm (adaptive threshold detection algorithm, ATDA) was developed with high precision of 97% [2]. Based on WSN and ATDA algorithm, Ding used a 3-node WSN to estimate the vehicle speed, and the precision was over 90% [11]. Similarly, Zhang developed a magnetic signature and length estimation algorithm to identify the vehicle type with binary proximity magnetic sensor networks and intelligent neuron classifier, and the simulations and on-road experiments obtained high recognition rate over 90% [12].

2.2. Traffic Data Quality Analysis. As the ITS applications expand in major urban areas, traffic data collection becomes more comprehensive. However, the quality of these traffic data is not as good as expected. They always contain many missing values or incorrect values and require careful "cleaning" to produce reliable results. The *bad data* and *missing data* have been an obstacle to ITS applications that use the data for performance.

Missing data is mainly caused by traffic data detectors failures or disruptions in communications. Given the continuous operation of most ITS traffic monitoring devices, *missing data* are almost inevitable. But they are relatively easy to be handled. For short periods of *missing data* identification, the method of time series analysis has been addressed. And the linear interpolation and neighborhood averages are natural methods to fill *missing data* with the data from the neighbors' data or the history [4].

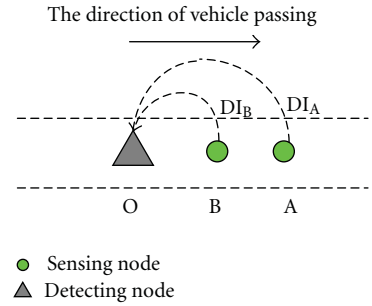


Figure 1: Topology of the traffic data detection model.

Compared with the processing of *missing data*, how to identify *bad data* and retrieve the *good data* from unreliable detectors are relatively complicated. In order to identify the *bad data*, there are many theories and practical methods proposed in the literature surveys, such as internal range checks, time series patterns, and historical patterns. For the single detector, Nihan introduced the conception of an acceptable region based on historical observations, which declared data to be inaccurate if they fell outside the region [13]. Similarly, Ki proposed an approach to check the error speed measurement with a fixed error-filtering algorithm [14]. While for the occasions of detectors extensively deployed in a large scale, the distributed methods coordinated with the nearby detectors should be proposed. C. Chen proposed a method to detect *bad data* with modeling the relationship between neighboring loops as linear and used linear regression to evaluate them [4]. Based on the similarity theory of traffic flow, Lelitha considered the conservation of traffic flow over a set of adjacent detectors to identify unreliable data [15]. Rajagopal presented a distributed and sequential algorithm for detecting multiple faults in a sensor network which worked by detecting the correlation statistics of neighboring sensors [16].

3. Modeling and Methodology

3.1. Traffic Data Detection and Aggregation

3.1.1. Traffic Data Detection by WSN. The model of distributed traffic data detection is shown as Figure 1, which is proposed by [11]. Compared with the other traffic data models based on WSN, it is more efficient with higher detection accuracy and lower energy consumption.

In the model, it is composed of two sensing nodes and one detecting node. The function of the sensing node is mainly to detect vehicle presence, whether a vehicle passing or not, and transmit the detection information (DI) as soon as a passing vehicle is detected. And the detecting node records the local time (T_A and T_B) as soon as the DIs of node A and node B is received. According the each timestamp

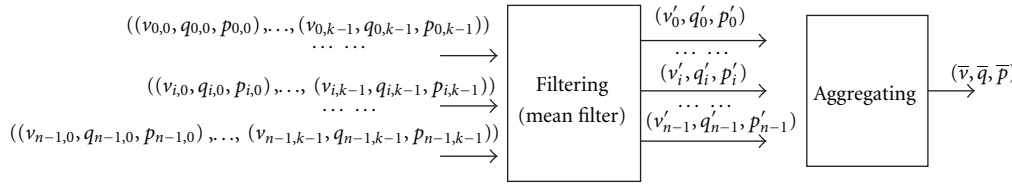


Figure 2: Aggregation model based on the mean filter.

(T_A and T_B), the corresponding values of speed, occupancy, and traffic flow can be calculated. Furthermore, the detecting node can control the working state of the two sensing nodes, to reduce the sensing nodes energy consumption.

Based on the traffic data detection model, the traffic data are detected as follows.

(a) Vehicle Passing Speed. The vehicle passing speed is calculated by

$$\begin{aligned} \Delta T &= T_A - T_B, \\ v' &= \frac{D_{AB}}{\Delta T}, \end{aligned} \tag{1}$$

where T_A and T_B are separately the time of the vehicle passing sensing nodes A and B in, respectively. D_{AB} is the distance between nodes A and B. v' is an instant value. While for more common, the vehicle speed v is the mean speed of vehicle crossing the sensing nodes during an observation interval, such as 5 minutes. The mean speed is calculated by (2).

$$v = \sum_{i=1}^{k} \frac{v'_i}{k}, \tag{2}$$

where k is the number of vehicles passed in the observation interval and can be set by the summation of accumulator.

(b) Occupancy. The occupancy p is the fraction of time when the detector is covered by vehicles in the observation interval. It can be calculated by

$$\begin{aligned} T_{\text{occ}} &= \sum_{i=1}^{k} \Delta T_i, \\ p &= \frac{T_{\text{occ}}}{T_{\text{segment}}}, \end{aligned} \tag{3}$$

where T_{segment} is the observation interval and k is the same as in (2).

(c) Traffic Flow. The traffic flow q is the number of vehicles crossing the traffic data detection model in an observation interval, and it is given by

$$q = k, \tag{4}$$

where k is the same as in (2).

3.1.2. Data Aggregation. Considering the distributed traffic data sources with a certain correlated character, the collected traffic data are correlated. Due to the correlation, the approach taking the correlation into account will outperform those which use the data directly.

In-network data aggregation is an important technique in WSN which exploits correlated sensing data and has been well studied in recent years [17].

For WSN that has irreplaceable batteries with limited energy capacity and poor data processing capability in practice, a distributed aggregation model based on the mean filter is proposed. The processing of the aggregation is shown as Figure 2.

Based on the distributed traffic data detection model, traffic data samples $((v_{i,k}, q_{i,k}, p_{i,k})$, $i = 0, 1, \ldots, n-1$, $k = 0, 1, \ldots, m-1)$ are calculated by the detecting node according to the DIs sent by n pairs of sensing nodes A and B. Here, m is the amount of vehicle passed in one observation interval. Finally, they are aggregated as $(\overline{v}, \overline{q}, \overline{p})$.

The aggregation model is as follows:

$$\begin{aligned} v'_i &= \frac{\left(\sum_{k=0}^{m-1} v_{i,k}\right)}{m}, \\ q'_i &= \frac{\left(\sum_{k=0}^{m-1} q_{i,k}\right)}{m}, \\ p'_i &= \frac{\left(\sum_{k=0}^{m-1} p_{i,k}\right)}{m}, \\ \overline{v} &= \frac{\left(\sum_{i=0}^{n-1} v'_i\right)}{n}, \\ \overline{q} &= \frac{\left(\sum_{i=0}^{n-1} q'_i\right)}{n}, \\ \overline{p} &= \frac{\left(\sum_{i=0}^{n-1} p'_i\right)}{n}. \end{aligned} \tag{5}$$

3.2. Model of Traffic Date Based on Cusp Catastrophe Theory. The existing models of traffic data mainly assume that the traffic data is at least locally gradual and linear, such as car following or hydrodynamic principle. However, according to recent researches, the discontinuity and catastrophe of traffic flow have been identified. But the previous traffic models are difficult to determine the breakpoint. The source of breakpoint is complex. Some breakpoints lie between the regime of free-flow condition and congested flow condition, and others are caused by the traffic incident.

The cusp catastrophe theory is used to describe the discontinuous phenomena of the natural. And researches have discussed the discontinuous behavior using the cusp catastrophe theory in traffic [18–20].

According to the basic cusp catastrophe [21, 22], the total potential energy (W) function of traffic data is

$$W(v) = av^4 + bqv^2 + cpv, \tag{6}$$

where v represents the state variable of W. q and p represent the control variable of W. And a, b and c are coefficients.

Based on the manifold function and the bifurcate equation of the cusp catastrophe, the model of traffic data based on the cusp catastrophe is defined as follows:

$$\begin{aligned} f(v,q,p) &= 4v^3 + 2mvq + np, \\ g(q,p) &= 8m^3q^3 + 27n^2p^2, \end{aligned} \tag{7}$$

where $m = b/a$; $n = c/a$.

In (7), m and n are coefficients, which can be captured by using the methods of mathematical statistics.

3.3. The Evaluation Function of Traffic Data Quality. Based on (7), the evaluation function of traffic data quality is defined as below:

$$\begin{aligned} h_1(f(v,q,p),g(q,p)) &= \|\bullet\| = \sqrt{f^2 + g^2}, \\ h_2(g(q,p)) &= ||g||, \end{aligned} \tag{8}$$

$(\overline{v}, \overline{q}, \overline{p})$ aggregated by the aggregation model is checked as the input of the evaluation function, and whether it is good or not, the data will be evaluated by (7). The conclusions of real-time traffic data validity are the following.

Conclusion 1. speed, flow, and occupancy are *good* when

$$h_1(f(v,q,p),g(q,p)) \le \varepsilon_{h1}. \tag{9}$$

Conclusion 2. speed is *error* , but flow and occupancy are *valid* , and the speed could be repaired when

$$\begin{aligned} &(h_1(f(v,q,p),g(q,p)) > \varepsilon_{h1}), \\ &(h_2(g(q,p) \le \varepsilon_{h2}). \end{aligned} \tag{10}$$

Conclusion 3. speed, flow, and occupancy are *bad* when

$$\begin{aligned} &(h_1(f(v,q,p),g(q,p)) > \varepsilon_{h1}), \\ &(h_2(g(q,p) > \varepsilon_{h2}), \end{aligned} \tag{11}$$

where ε_{h1} and ε_{h2} are the corresponding threshold level.

In Conclusion 2, v can be repaired by q and p, where $f(v,q,p) = 4v^3 + 2mvq + np = 0$. In this equation, according to Cardano's formula, **v** is solved. The value **v** could be more than one solution, and the one is selected which is the most closest to the value in previous interval.

4. DTDDQA Algorithm

In this section, we design a distributed traffic data detection and quality analysis algorithm, named *distributed traffic data detection and quality analysis* (DTDDQA) algorithm, which consists of two steps. The first step is the traffic data collection and aggregation which aggregates the traffic data collected from n nodes in every 5 minutes. The second step is the real-time data quality analysis according to the data profiling of traffic data constructed by the model based on cusp catastrophe theory. As the output, the *good data* can be verified and some *bad data* can be repaired.

Based on DTDDQA algorithm, the processing of traffic data analysis is shown as Figure 3. The whole processing is running on the AP node. T_{Ai} and T_{Bi} sent by the sensor nodes A_i and B_i, when vehicles passed. As the input of DTDDQA algorithm, they are firstly calculated to (v_i, q_i, p_i) by (2) (3), and (4). Then, they are aggregated to $(\overline{v}, \overline{q}, \overline{p})$ with other sensor nodes every 5 minutes. Finally, $(\overline{v}, \overline{q}, \overline{p})$ is checked by data quality analysis, and the *good data* and the *repaired data* stored as (v, q, p) are outputted.

Based on the method mentioned above, a state machine is designed to perform the processing of DTDDQA algorithm adaptively for traffic data detection and quality analysis, as shown in Figure 4. The T_{Ai} and T_{Bi} are the input, and the *good data* are the output. The transition states in state machine have several steps as follows.

(1) *Initialization state*. It is mainly to set the initial value of system (m, n, ε_{h1}, and ε_{h2}) and then transit to *Distributed detection state*.

(2) *Distributed detection state*. In this state, (v_i, q_i, p_i) is calculated according T_{Ai}, and T_{Bi} sent by the sensor nodes A_i and B_i and then transit to *Aggregation state*.

(3) *Aggregation state*. It is aggregated to $(\overline{v}, \overline{q}, \overline{p})$ every 5 minutes and then transit to *Analysis state*.

(4) *Analysis state*. It uses cusp catastrophe theory to evaluate $(\overline{v}, \overline{q}, \overline{p})$. If it is evaluated as *good*, it will transit to *Output state*. While the data is invalid, it will roll back to *distributed detection state* in the case of Conclusion 3, or transit to *repair state* in the case of Conclusion 2.

(5) *Output state*. Store *good data* and *repaired data,* or transmit these data to the data center, and then transit to *distributed detection state*.

(6) *Repair state*. According to (7), the average speed is rectified by $\overline{q}$ and $\overline{p}$. Then, it will transit to *distributed detection state*.

As a result, the state machine can be running automatically without the end state. In actual operations, the system needs to be stopped or restarted manually at a pinch.

5. Simulation Results and Analysis

The presented traffic data collection and data quality analysis are investigated in this paper via simulation with VISSIM which is a professional traffic simulation software.

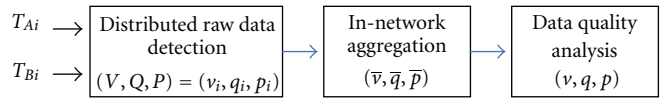


Figure 3: Processing of traffic data quality analysis based on the DTDDQA algorithm.

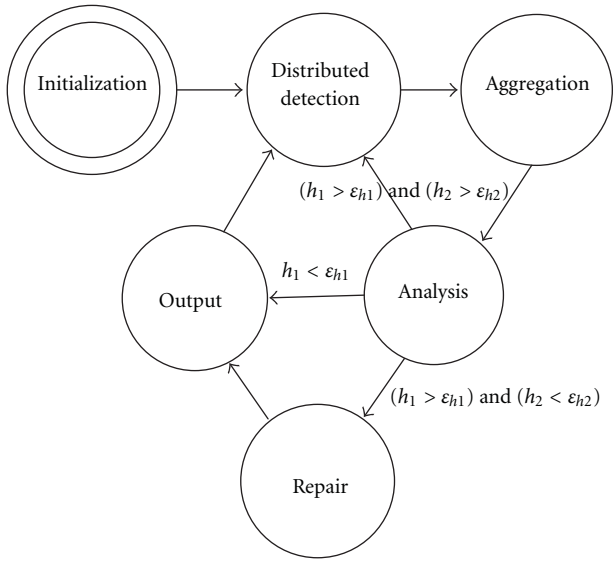


Figure 4: State machine for the DTDDQA algorithm.

5.1. Simulation Setup. The traffic data set adopted in this paper comes from the Dalian Transportation Management Center. It has been collected by the inductive loop detectors deployed in the Huanghe Road, Dalian, China. And the data set contains the traffic data (speed, flow, and occupancy) from 0:00 AM to 12:00 PM, December 2, 2008.

The Huanghe Road is a road with four-lane in each direction. The traffic flow is quite complicated, and the traffic state varies obviously.

Based on the traffic data set, we reinstate the traffic scene using VISSIM. In the simulation, a distributed traffic data collection model based on WSN is deployed on the westbound direction's entrance of Huanghe Road, as shown in Figure 5.

The traffic data collection model is deployed at point A in Figure 5(a). The model contains four observation points (OP) and one AP node. Each OP is composed of 3 sensor nodes and deployed in each lane. The functions of each sensor nodes are described in the Section 3.1. The AP node collects the traffic flow data from OPs and performs the algorithm DTDDQA to analyze and clean the traffic data.

5.2. Performance of Traffic Data Detection and Aggregation. In this subsection, we focus on the traffic data collection based on WSN and study the characters of traffic flow, especially whether the traffic flow has correlation and catastrophe. Furthermore, we analyze the performance of the aggregation of distributed collection model based on WSN.

Using the experimental data collected with the platform as above, the statistics of average speed, average occupancy rate, and flow from OP1 to OP4 in every 5 minutes of 24-hours are shown in Figure 6.

It is obvious that the traffic flow is catastrophic and nonlinear.

For the performance of the aggregation and the correlation of distributed collection model based on WSN, both standard deviation and relative performance indices are utilized in this paper.

The standard deviation index of ith OP (SD_i) is defined as

$$\mathrm{SD}_i = \sqrt{\frac{\sum_{j=0}^{M-1} \left(x'_{i,j} - \overline{x}_j\right)^2}{M}}. \tag{12}$$

According to [23], the corresponding relative performance index of ith OP (PI_i) is defined as

$$\mathrm{PI}_i = \frac{\mathrm{SD}_i}{1/M \sum_{j=0}^{M} x'_{i,j}}, \tag{13}$$

where $x'_{i,j}$ is the jth value of ith OP, x could be the value of v, p, or q. $i = 0, 1, \ldots, N-1, N$ is the number of OP, here $N = 4$; $j = 0, 1, \ldots, M-1$, M is the amount of sample, here $M = 288$, and $\overline{x}_j$ is the jth value of aggregation processing. The better corresponding relative performance, the smaller value of PI_i.

So, the performance of the aggregation of distributed collection model is shown as Table 1 in which the SD and PI between the value of aggregation processing and the collection of OP every 5 minutes are analyzed.

It is obvious that the collection data of each OP is certainly correlated. While the SD and PI of OP4's values to the aggregation values are largest, it is mainly because that the lane of OP4 is the right-turn entrance of the upper intersection and the traffic flow is a slightly different from that the other three lanes.

5.3. Performance of Traffic Data Quality Analysis. In this subsection, we focus on performance of the DTDDQA algorithm. Without loss of generality, we compare the performance of the following two traffic data quality analysis algorithms:

(1) the DTDDQA algorithm,

(2) the *error data* identification method in [14].

We analyze these two algorithms based on the data set achieved from loop sensor. Firstly, the data set is aggregated by (5), and then, we analyze the traffic data quality using these two algorithms based on the aggregated data set.

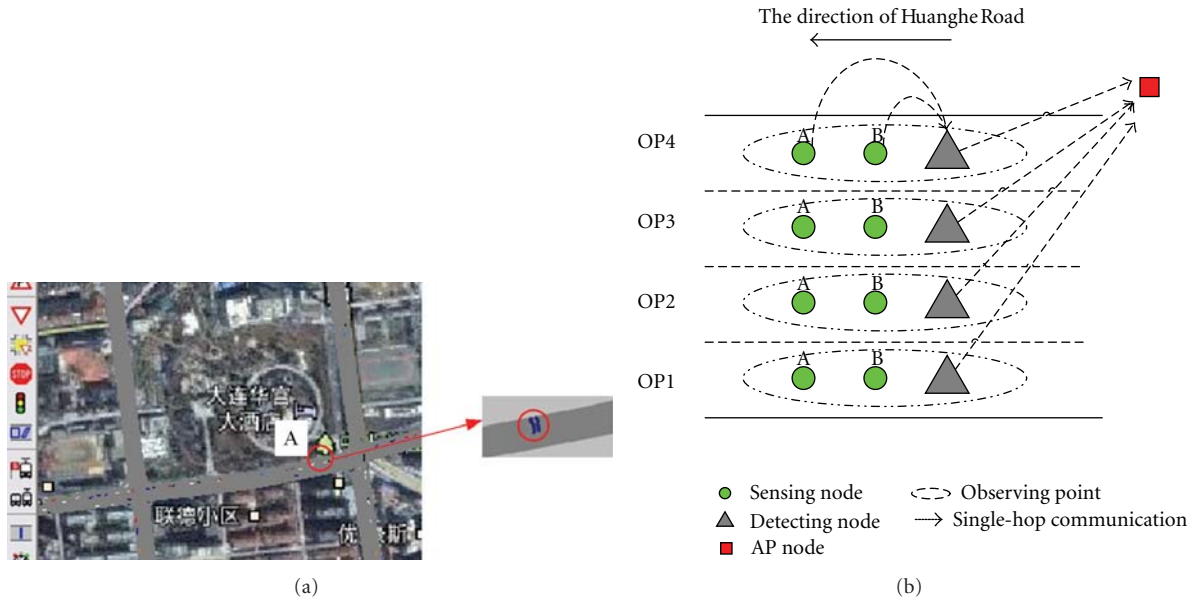


Figure 5: Traffic scene reinstated by VISSIM: (a) The test bed of WSN in VISSIM; (b) The layout of WSN.

Table 1: Performance of the aggregation.

	Speed		Flow		Occupancy	
	SD (km/h)	PI (%)	SD (veh/5 min)	PI (%)	SD (%)	PI (%)
OP1	4.490278	9.02	3.62934	15.3	1.594184	12.89
OP2	2.760417	7.10	6.164063	19.2	1.653906	15.07
OP3	6.361632	10.21	10.42795	17.95	5.290885	24.09
OP4	11.99531	19.54	11.93663	29.2	7.017274	32.37

In order to evaluate the algorithms, 40 of 288 in the aggregated data set are modified to *bad data* manually. Among the 40 *bad data*, 20 samples are only the speed value altered (regarded as *repairable data*) and in the other 20 samples' speed and flow are altered (regarded as *error data*). The new data set will be analyzed separately using these two algorithms. The results illustrate that the performance of DTDDQA algorithm is better. The performance detailed is as follows.

5.3.1. Performance of DTDDQA. The new data set is analyzed by DTDDQA algorithm, and the result is shown in Figure 7. The samples of the aggregated data set are input, and three integer values predefined (1, 2 and 3) are output. Based on the three *Conclusions* of (8), 1 represents that the sample is *good*; 2 is stated that the sample is *repairable* (occupancy and flow are exact, but velocity is error); 3 means that this sample is *error* and *irreparable*. The performance analysis result is shown in Table 2. The identification of *good data* is about 87.71%, and it increases to 94.40% if the *repaired data* are included. And the identification of *bad data* is 75.00%.

5.3.2. Performance of the Data Quality Algorithm in [14]. The existing traffic data quality analysis algorithms mainly focus on the data collected from freeway. However, the traffic flow of freeway tends to linear and gradual. In order to study whether the algorithms are fit for the traffic data in urban roadway, the algorithm in [14] is tested. It is proposed with a filter to process traffic data, which is only dealt with the speed value. When the data deviation is more than 5%, the sample will be removed.

Table 2: DTDDQA results.

	Good data	Repairable data	Error data
Actual data set	248	20	20
Output data set	220	33	25

Using this algorithm to analyze the aggregated data set, the result is shown in Figure 8. Similarly, the speed values of the aggregated data set are the input, and two integer values predefined (0 and 1) are output, where 0 means the data is *good* and 1 represents the data is *error*. As a result, there are 176 samples out of 288 considered to be *good data*, so the identification rate is 70.97%, which shows that the performance of the algorithm to the data from the urban roadway is not as good as it does from the freeways.

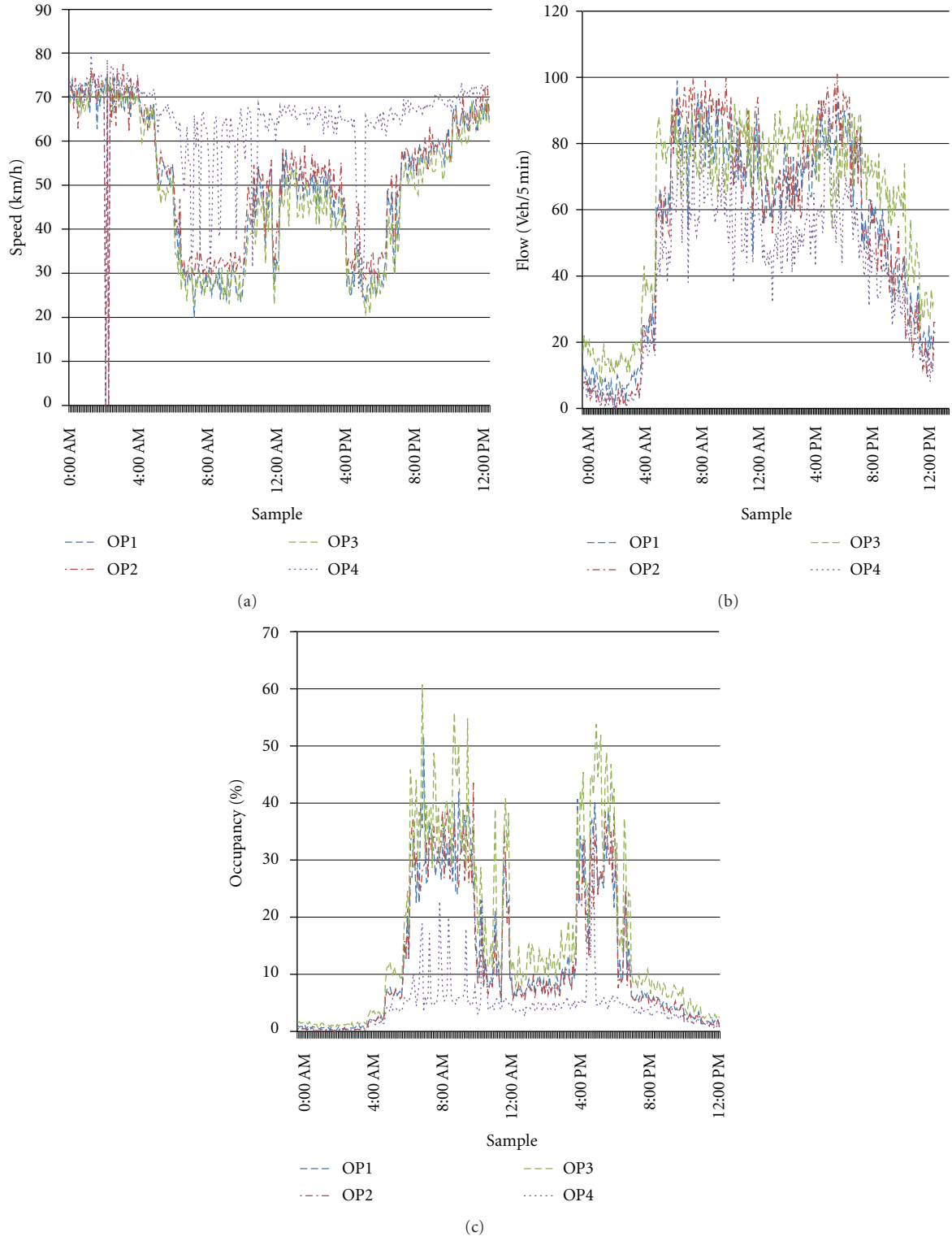


Figure 6: 24-hours traffic data samples collected from OP1 to OP4 in every 5 minutes: (a) Speed; (b) Flow; (c) Occupancy.

6. Conclusions

WSN is a revolution in applications of information sensing and collection, and consequently, it has broad prospect in the ITS. In this paper, we develop a distributed algorithm (DTDDQA) for urban traffic data detection and quality analysis based on WSN. In this algorithm, we firstly propose an aggregation model based on the mean filter to process the distributed data samples collected by WSN and then present an evaluation equation and data quality analysis

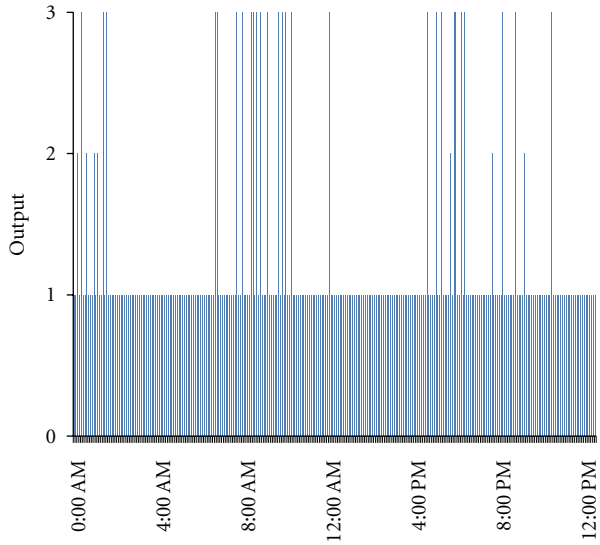


Figure 7: The traffic data quality analysis results with DTDDQA.

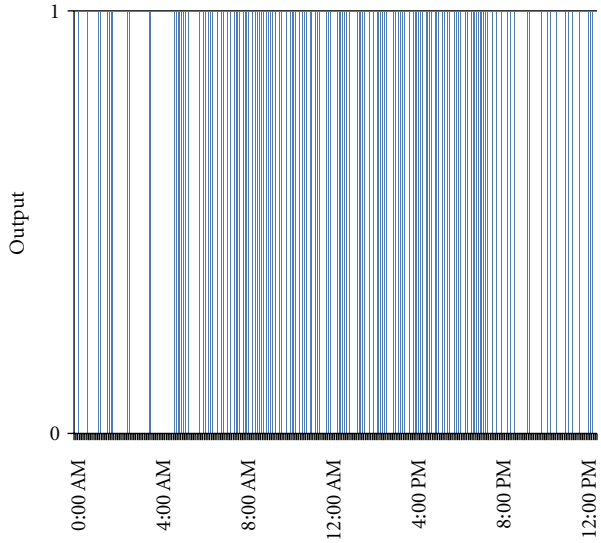


Figure 8: The traffic data quality analysis results with the algorithm in [14].

model based on the cusp catastrophe theory to identify the bad data and try to repair them. A number of simulations are conducted based on the real data samples collected from on-road detectors. As a result, with the processing of data quality analysis and data recovery, this algorithm improves the correctness and robustness of traffic data collection using WSN.

Although we focused on the effort in developing a general algorithm for urban traffic data detection and quality analysis, the difference and variation of traffic flow characters in the live detection scenario should be taken into account. For future work, we propose to study the algorithm further to be more self-adaptable and self-adjustable according to the traffic data detection based on the active-learning mechanism.

Acknowledgments

This work was supported in part by the Key Program of National Science Foundation of China under Grant no. 60873256. The authors would like to thank the Dalian Department of Transportation for providing the utilized real traffic measurement data and the anonymous reviewers for their valuable comments.

References

[1] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century field experiment," *Transportation Research C*, vol. 18, no. 4, pp. 568–583, 2010.

[2] S. Cheung and P. Varaiya, *The Quality of Loop Data and the Health of California's Freeway Loop Detectors*, PeMS Development Group, University of California, Berkeley, Calif, USA, 2002.

[3] P. Wu, *Automated Data Collection, Analysis, and Archival (Final Report)*, University of Utah, Salt Lake City, Utah, USA, 2003.

[4] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems," *Transportation Research Record*, no. 1855, pp. 160–167, 2003.

[5] S. Cheung and P. Varaiya, *Traffic Surveillance by Wireless Sensor Networks: Final Report*, University of California, Berkeley, Calif, USA, 2007.

[6] M. Tubaishat, P. Zhuang, Q. Qi, and S. Yi, "Wireless sensor networks in intelligent transportation systems," *Wireless Communications and Mobile Computing*, vol. 9, no. 3, pp. 287–302, 2009.

[7] H. Kowshik, D. Caveney, and P. R. Kumar, "Safety and liveness in intelligent intersections," *Hybrid Systems: Computation and Control*, vol. 4981, pp. 301–315, 2008.

[8] H. Amine, K. Robert, and P. Varaiya, "Wireless magnetic sensors for traffic surveillance," *Transportation Research C*, vol. 16, no. 3, pp. 294–306, 2008.

[9] J. Jeahoon, G. Shuo, and H. Tian, "APL: autonomous passive localization for wireless sensors deployed in road networks," in *Proceedings of the 27th IEEE Communications Society Conference on Computer Communications (INFOCOM '08)*, pp. 583–591, Phoenix, Ariz, USA, April 2008.

[10] S. Y. Cheung, S. Coleri, B. Dundar, S. Ganesh, C. W. Tan, and P. Varaiya, "Traffic measurement and vehicle classification with single magnetic sensor," *Transportation Research Record*, no. 1917, pp. 173–181, 2005.

[11] N. Ding, G. Tan, H. Ma, M. Lin, and Y. Shang, "Low-power vehicle speed estimation algorithm based on WSN," in *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems (ITSC '08)*, pp. 1015–1020, Beijing, China, December 2008.

[12] W. Zhang, G. Z. Tan, H. M. Shi, and M. W. Lin, "A distributed threshold algorithm for vehicle classification based on binary proximity sensors and intelligent neuron classifier," *Journal of Information Science and Engineering*, vol. 26, no. 3, pp. 769–783, 2010.

[13] N. L. Nihan, X. Zhang, and Y. Wang, "Detector data validity," Tech. Rep., Washington State Transportation Center, 1990.

[14] Y. K. Ki and D. K. Baik, "Model for accurate speed measurement using double-loop detectors," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 4, pp. 1094–1101, 2006.

[15] V. Lelitha and L. R. Rilett, "Loop detector data diagnostics based on conservation-of-vehicles principle," *Transportation Research Record*, no. 1870, pp. 162–169, 2004.

[16] R. Rajagopal, X. L. Nguyen, S. C. Ergen, and P. Varaiya, "Distributed online simultaneous fault detection for multiple sensors," in *Proceedings of the International Conference on Information Processing in Sensor Networks (IPSN '08)*, pp. 133–144, St. Louis, Mo, USA, April 2008.

[17] S. Ozdemir and Y. Xiao, "Secure data aggregation in wireless sensor networks: a comprehensive overview," *Computer Networks*, vol. 53, no. 12, pp. 2022–2037, 2009.

[18] X. Lignos, G. Ioannidis, and A. N. Kounadis, "Non-linear buckling of simple models with tilted cusp catastrophe," *International Journal of Non-Linear Mechanics*, vol. 38, no. 8, pp. 1163–1172, 2003.

[19] J. Guo, X. L. Chen, and H. Z. Jin, "Research on model of traffic flow based on cusp catastrophe," *Control and Decision*, vol. 23, no. 2, pp. 237–240, 2008.

[20] S. Kang and R. Jayakrishnan, *Theoretical Analysis of Catastrophes in Traffic*, University of California, Irvine, Calif, USA, 1998.

[21] F. P. Navin, "Traffic congestion catastrophes," *Transportation Planning and Technology*, vol. 11, no. 1, pp. 19–25, 1986.

[22] Y. Zhang and Y. Pei, "The application of cusp catastrophe theory in traffic flow prediction," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pp. 628–631, 2003.

[23] Y. Wang, M. Papageorgiou, and A. Messmer, "Real-time freeway traffic state estimation based on extended Kalman filter: a case study," *Transportation Science*, vol. 41, no. 2, pp. 167–181, 2007.