

**ACTIVE SECURITY MECHANISMS FOR WIRELESS SENSOR  
NETWORKS AND ENERGY OPTIMIZATION FOR PASSIVE SECURITY  
ROUTING**

A Dissertation

Submitted to the Graduate Faculty of the  
Louisiana State University and  
Agricultural and Mechanical College  
in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

in

The Department of Computer Science

by

Lydia Ray

B.Sc, University of Calcutta, 1996

M.Stat., Indian Statistical Institute, 1998

August 2005

# Acknowledgements

First, I wish to express my sincere gratitude to my major professor Dr. Rajgopal Kannan. Without his guidance, inspiration and enthusiasm this dissertation would not become a reality.

My gratitude extends to my department chair Dr. S.S. Iyengar, my committee members Dr. Arjan Durrezi and Dr. Ram Vaidynathan for their advice and cooperation.

I wish to thank all other professors in my department, especially Dr. Donald Kraft, Dr. Jinhua Chen, Dr. S. Kundu and Ms. Savita Pinepalli for their guidance and advice throughout my coursework.

I am highly grateful to my professors in Indian Statistical Institute, especially Dr. R.B. Bapat and Dr. S.R. Mohan whose help and encouragement enabled me to pursue my doctoral research.

I would like to express my gratitude to my friends, Anuradha Krishnan and Debapriya Sen, who have constantly supported and encouraged me to continue and complete my doctoral study. I also thank to my fellow researcher Ramaraju Kalindindi for his help and several stimulating discussions about research.

I wish to thank Lynette Jackson and Vera Watkins for their help and support.

# Table of Contents

<b>Acknowledgements</b> . . . . .	<b>ii</b>
<b>Abstract</b> . . . . .	<b>v</b>
<b>Chapter 1: Introduction</b> . . . . .	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Thesis Objectives . . . . .	2
1.3 Thesis Contributions . . . . .	3
1.4 Thesis Organization . . . . .	5
<b>Chapter 2: Overview of Sensor Networks</b> . . . . .	<b>6</b>
2.1 Introduction . . . . .	6
2.2 Applications of Sensor Networks . . . . .	7
2.3 Design Factors and Communication Architecture of Sensor Networks . . . . .	9
2.4 Protocol Stack . . . . .	11
2.5 Conclusion . . . . .	16
<b>Chapter 3: Bootstrapping Problem and Key Pre-distribution</b> . . . . .	<b>17</b>
3.1 Introduction . . . . .	17
3.2 Bootstrapping Problem in a Wireless Sensor Network . . . . .	19
3.3 Evaluation of Different Key Agreement Schemes with Respect to Sensor Networks . . . . .	20
3.4 Literature Review . . . . .	25
3.5 Two-Phase Key Pre-distribution Mechanism . . . . .	29
3.6 Metrics for Measuring Security-Performance Tradeoffs . . . . .	30
3.7 Secure Network Connectivity: Analytical Results . . . . .	32

3.8	Network Resiliency against Enemy Attack: Analytical Results . . . . .	33
3.9	Simulation Results . . . . .	42
3.10	Implementation Issue: Creating Sorted Shared Key Lists . . . . .	47
3.11	Conclusion . . . . .	47
<b>Chapter 4: Secure Data Aggregation . . . . .</b>		<b>48</b>
4.1	Introduction . . . . .	48
4.2	Literature Review on Data Aggregation and Security . . . . .	49
4.3	Attack Model and Security Goals . . . . .	52
4.4	Problem Setup and Statement . . . . .	53
4.5	Weighted Data Aggregation . . . . .	54
4.6	Error Analysis . . . . .	57
4.7	Selection of Parameters . . . . .	59
4.8	Performance Evaluation . . . . .	60
4.9	Compromised Aggregators: Problem and Solutions . . . . .	63
4.10	Solutions: Our Approach . . . . .	64
4.11	Conclusion . . . . .	69
<b>Chapter 5: Energy Optimized Routing for Passive Security . . . . .</b>		<b>70</b>
5.1	Introduction . . . . .	70
5.2	Related Works . . . . .	73
5.3	Analytical Model of Length-Energy-Constrained Routing . . . . .	74
5.4	Distributed Protocol Implementation . . . . .	78
5.5	Selection of Parameters . . . . .	84
5.6	Performance Evaluation . . . . .	86
5.7	Conclusion . . . . .	91
<b>Chapter 6: Summary of Thesis and Future Work . . . . .</b>		<b>92</b>
6.1	Summary of Thesis . . . . .	92
6.2	Future Work . . . . .	94
<b>Bibliography . . . . .</b>		<b>96</b>
<b>Vita . . . . .</b>		<b>104</b>

# Abstract

Wireless sensor networks consisting of numerous tiny low power autonomous sensor nodes provide us with the remarkable ability to remotely view and interact with the previously unobservable physical world. However, incorporating computation intensive security measures in sensor networks with limited resources is a challenging research issue. The objective of our thesis is to explore different security aspects of sensor networks and provide novel solutions for significant problems.

We classify security mechanisms into two categories - active category and passive category. The problem of providing a secure communication infrastructure require active security measurements. Key pre-distribution is a well-known technique in this class. We propose a novel 2-Phase technique for key predistribution based on a combination of inherited and random key assignments from the given key pool to individual sensor nodes. We develop an analytical framework for measuring security-performance tradeoffs of different key distribution schemes. Using rigorous mathematical analysis and detailed simulation, we show that the proposed scheme outperforms the existing solution in every performance aspect.

Secure data aggregation in wireless sensor networks is another challenging problem requiring active measures. We address the problem of stealthy attack where a compromised node sends wrong/fictitious data as a reply to a query. We propose a novel probabilistic accuracy model which enables an aggregator to compute accuracy of each sensor reading by exploiting spatial correlation among data values. We also propose some novel, energy efficient statistical methods to enable a user accept the correct value with a high probability.

Increasing network lifetime is a passive security mechanism which enables many security mechanisms to work more efficiently. We define length-energy-constrained optimality criteria for energy-optimized routes that impose uniform energy distribution across the network, thus preventing expedited network partition. We propose three different distributed, nearly-stateless and energy efficient routing protocols that dynamically find optimal routes and balance energy consumption across the network. We show that global energy information acquired through this process utilized in conjunction with energy depletion control in the sensornet ensures a significant improvement in terms of network lifetime.

# Chapter 1

## Introduction

### 1.1 Motivation

Recent advances in wireless communication and remote sensing technology have led to the emergence of wireless sensor networks as a new tool for closely observing the world around and within us. Several unique features have distinguished sensor networks from other traditional wireless networks and made them attractive for a broad spectrum of applications.

- Sensor nodes being tiny, low power and low capacity, a sensor network can be deployed on a very large scale (several orders of magnitude higher than traditional wireless networks) in resource limited and harsh environments where human intervention is not possible or desirable [3], [42]. For example, battlefields, habit monitoring in forests, ecological contamination sites, seismic zones and so on.
- Nodes being equipped with integrated sensing and data processing capabilities, sensor networks can acquire spatio-temporally dense and localized data at a very low expense and frequently in time [39, 15, 28].
- Sensors being able to be very close to the phenomena of interest, distributed sensing improves signal-to-noise ratio and allows greater fidelity of data with the potential to reveal previously unobservable phenomena in the physical world.

Interestingly however, the strengths of a wireless sensor network also contributes to its main drawbacks. In order to make large scale and low expense deployments possible, sensor nodes are made low in capacity and tiny with limited hardware facilities. Sensor fields being deployed in hostile environments, nodes are unattended and unreplenishable. These severe operational constraints expose sensor networks to the risk of a variety of security

threats and attacks. Different types of attacks include denial of service, spoofed, altered, or replayed routing information, selective forwarding, stealthy attack and so on. Sensor network researchers, therefore, face the challenge of maximizing network security subject to stringent resource constraints [46, 50, 55, 54, 49]. Apart from resource constraints, the following characteristics of sensor networks add to the challenge of security research.

- The range of application domains varies widely and security threats for sensor networks are extremely application specific. For example, battlefield applications are more prone to denial of service attacks while habitat monitoring or other environmental applications are subject to eavesdropping or stealthy attack. Therefore, there exists no single unique solution that can provide the most heightened security for all types of sensornets at the least expense of resources.
- In-network data aggregation makes the design of a secure routing protocol more complicated. Intermediate routers in conventional networks do not have access to the content of messages, thereby enabling end-to-end security mechanisms such as SSH or SSL possible. A secure routing protocol only requires to guarantee message availability. However, in sensor networks, intermediate nodes need direct access to the content of the messages in order to perform in-network data processing. Thus end-to-end security mechanisms are not feasible.
- Uneven energy consumption across sensor networks leads to early network partition even when most of the nodes are active, resulting in a drastic reduction in the integrity of the network component still attached to the base station. A smaller network size will enable an adversary to perform more expensive attacks at lesser cost.

## 1.2 Thesis Objectives

We classify security mechanisms for a wireless sensor network into two broad categories:

- active security mechanisms
- passive security mechanisms.

While different protocols for encrypting data, building a secure communication infrastructure or preventing specific attacks belong to the active category, resource efficient protocols and algorithms for different activities at different layers of protocol stack are classified as passive security mechanisms. Active security mechanisms require intense computations but are often restricted by resource constraints of sensor networks. Passive security mechanisms

aim at saving resource energy and increasing network lifetime as much as possible, thereby preventing more powerful attacks by an adversary and enabling implementation of more computation-intense security mechanisms that provide more tightened security.

In this thesis, we focus on the following security problems:

- *bootstrapping problem* and *secure data aggregation* in the active mechanism category,
- *energy optimized routing* in the passive mechanism category.

Objectives of the thesis are as follows:

- To provide an overview of sensor networks that enables in-depth understanding of the security problems dealt in the thesis.
- To provide thorough survey of related previous works along with their drawbacks.
- To provide new solutions that outperform the existing solutions by overcoming the corresponding drawbacks.

### 1.3 Thesis Contributions

There are four main contributions of this thesis:

- The thesis contains a comprehensive overview of wireless sensor networks and a survey of previous works along with their corresponding drawbacks in the fields considered in the thesis.
- We consider the bootstrapping problem in wireless sensor networks and have following contributions:
  - We propose a novel key pre-distribution algorithm named *2-Phase key pre-distribution* that improves the connectivity and capture-resiliency properties of loading sensor nodes with a combination of *randomly derived* and *inherited* keys.
  - We develop novel quantitative metrics that measure any key predistribution schemes' security-performance tradeoffs in terms of the network resiliency to node/key capture, the number of available secure links and the key (memory) requirement per node for a given level of connectivity.
  - We analytically evaluate our algorithm using the proposed framework and compare it with the existing algorithm named *random key pre-distribution*.

- We show both analytically and by simulation that our solution outperforms the existing random pre-distribution scheme by simultaneously providing better connectivity and higher resiliency to enemy attack.
- We address the problem of enabling secure information aggregation and have the following contributions:
  - Previous works on this problem provides a solution by unrealistically limiting the number nodes that an adversary can compromise. We provide a more general solution which does not require this assumption.
  - We provide an analytical model of computing data accuracy based on *spatial correlation* of data values.
  - We provide a novel algorithm which computes a *weighted aggregate* of data by attaching less weights to the sensor readings which are more likely to be wrong according to the proposed data accuracy model.
  - We show by simulation that the weighted aggregate is more accurate than a simple aggregate.
  - We consider the problem of data aggregation when an aggregator is corrupted. We propose several solutions based on statistical estimation of parameters in a multi-aggregator set-up.
- We address the problem of length-energy-constrained routing in a wireless sensor network to increase network lifetime as a passive security measure and have the following contributions:
  - We propose a decision theoretic paradigm for solving the problem of finding energy-optimal routing paths with bounded path length. We define a routing game in which sensors obtain benefits by linking to healthy nodes while paying a portion of path length Thus sensor nodes modelled as intelligent agents cooperate to find optimal routes.
  - We show that using limited global information on node-energies combined with local geographic forwarding can remarkably improve network lifetime. We propose three distributed and nearly stateless inter-cluster routing protocols which consider path length and energy costs and yield significant improvement in network lifetime.
  - We evaluate our routing protocols using the ns-2 simulator. Simulation results indicate that all the protocols are enormously effective in reducing energy devi-

ation, thereby leading to equitable residual energy distribution across the sensor network. Thus the protocols should have a significant impact on sensor network survivability.

## 1.4 Thesis Organization

The thesis is organized as follows. Chapter 2 provides an overview of wireless sensor networks and attack models together with reviews of related literatures. In chapter 3, we deal with the bootstrapping problem. We provide literature reviews and our solution together with simulation results. In chapter 4, we present the problem and solution of secure data aggregation together with relevant literature reviews. In chapter 5, we consider the problem of energy-optimized routing and provide our solutions together with an overview of related works. Lastly, chapter 6 contains a summery of the thesis and discussion of future works.

# Chapter 2

## Overview of Sensor Networks

### 2.1 Introduction

The explosive growth of wireless communications and electronics has enabled the development of low-cost, low-power, multifunctional sensors. Composing these sensor nodes into sophisticated computational and communication infrastructures to form wireless sensor networks is a unique challenge to modern technology.

Wireless sensor networks consist of numerous tiny low-cost, low-power sensor nodes deployed densely in terrains where human intervention is impossible [1, 2, 3]. Individually, each node is autonomous and has *short range*. Collectively, these nodes cooperate to collect and disseminate information from a large area. The main features of a sensor network are as follows:

- The position of sensor nodes need not be engineered or predetermined, thereby enabling random deployment of nodes in inaccessible areas and disaster relief operations.
- Wireless sensor networks possess self-organizing capabilities.
- Sensor nodes being fitted with an on-board processor can locally carry out simple computations and transmit only the required and partially processed data.

These features enable a wireless sensor network to lend itself to a very different set of applications than a network with a small number of powerful long-range sensors. A centralized, long-range approach fail in complex, cluttered environments where line-of-sight paths are typically very short. For example, satellites are not very good at detecting individual animals in a forest, objects in a building, or chemicals in the soil. Moving to a distributed

collection of shorter-range sensors can dramatically reduce the effect of clutter. By increasing the number of vantage points, it is more likely that an area will be viewable, even when line-of-sight paths are short. Many interesting phenomena which are localized cannot be effectively sensed from a long distance. Temperature and humidity are excellent examples. In such cases, distributed sensing improves signal-to-noise ratio, thereby allowing greater fidelity in the collected data.

This chapter is organized as follows. In the next section, we present some important applications of sensor networks. Then we provide an overview of design factors and wireless sensor network architectures. In the next section, a brief overview of different layers of sensor network protocol stack has been provided along with some important protocols for each layer.

## **2.2 Applications of Sensor Networks**

### **2.2.1 Military and Defense Applications**

The main battlefield applications of sensor networks include surveillance and target tracking, real time monitoring, reporting threat data and giving precise location information. Defense Advanced Research Projects Agency (DARPA) has started Sensor Information Technology Program (SenseIT) [4] which have been pursuing research in battle field oriented application of sensor networks. SenseIT networks support low latency, energy-efficient operation, built-in autonomy and survivability, and low probability of detection of operation, thereby enabling important battlefield activities such as detection, identification, and tracking of threats, as well as targeting and communication, both within the network and to outside the network.

Tactical Automated Security System (TASS) [5] is used for Perimeter security.

The Wide-Area Tracking System (WATS) [6] is a network of gamma and neutron detectors and communications links which can detect and track any ground -delivered nuclear material. The sensors can be permanently deployed at chosen locations or mounted in vans for deployment on demand to protect specific areas for specific situations or events.

Another military application is Laboratory-developed Counterproliferation Analysis and Planning System (CAPS) [6] which can model the various processes (chemical, biological, metallurgical) used by proliferators to build weapons of mass destruction and their delivery systems. Thus, CAPS helps users identify critical processing steps or production facilities whose destruction can prevent that country from producing weapons of mass destruction.

### 2.2.2 Habitat Monitoring

Wireless sensor networks being capable of collecting localized data from complex and cluttered environment have widespread applications for habitat sensing for biocomplexity mapping. One of the applications is Great Duck Island System (GDI), developed in Great Duck Island, Maine by researchers from UCB/Intel Research Laboratory, to monitor the behavior of storm petrel [7]. Sensors used in this application are Atmel Atmega running at 4 MHz, 916 MHz radio to provide bidirectional communication at 40 kbps and a pair of AA batteries to provide energy.

PODS [9] is a remote ecological micro-sensor network which investigate why endangered species of plants will grow in one area but not in neighboring areas. This network uses nodes called Pods that consist a computer, radio transceiver and environmental sensors sometimes including a high resolution digital camera and relay sensor data via wireless link back to the Internet. PODS collect two types of sensor data - weather data that are collected every ten minutes and image data that are collected once per hour.

Some other methods for habitat monitoring are target classification by maximum cross-correlation between measured acoustic signal and reference signal, localization using TDOA-based beamforming, and data reduction using zero-crossing rate technique [8]. A prototype testbed consisting of iPAQs is used to evaluate the performance of those target classification and localization methods.

### 2.2.3 Environmental Observation and Forecasting System

Environmental Observation and forecasting system (EOFS) is defined as a large distributed system which spans a large geographic area to forecast phenomena such as flood, draught, environmental pollution etc. Automated Local Evaluation in Real-Time (ALERT) [10] is the first well-known wireless sensor networks deployed in real world. Developed by the National Weather Service in 4 the 1970's, ALERT provides real-time data on rainfall and water level to evaluate the possibility of flooding. ALERT uses meteorological/ hydrological sensors, such as water level sensors, temperature sensors, wind sensors etc which transmit data via light-of-sight radio communication from the sensor site to the base station. A Flood Forecast Model processes those data and issues automatic warning. Currently ALERT is deployed across most of the western United States.

### 2.2.4 Health Applications

Applications in this category are designed for telemonitoring of human physiological data. Promising research works are being conducted on projects of embedding wireless biomedical

chips inside human body to meet additional safety and reliability challenges. Smart Sensors and Integrated Microsystems(SSIM) [11] is one such example that builds retina prosthesis chips consisting of 100 microsensors that can be implanted within human eye allowing patients with limited or no vision to see at an acceptable level. The wireless communication system in this application provides feedback control, image identification and validation using a LEACH [18] like cluster-based system or a tree-based system. Some similar applications are Glucose level monitors, Organ monitors, Cancer detectors and General health monitors.

Realization of the above mentioned applications requires some special design features which distinguish a wireless sensor network from traditional wireless ad hoc networks. In the next subsection, some of the special design issues of wireless sensor networks are discussed.

## 2.3 Design Factors and Communication Architecture of Sensor Networks

Sensor Networks are generally scattered in a sensor field. the network. Sensor nodes collect information according to the query, perform in-network data processing and send that information back to the sink through a reverse multicast tree. The sink may communicate with the *task manager* node via internet or satellite. We provide brief descriptions of different components of a sensor network as follows.

### 2.3.1 Hardware Design of Sensor Nodes

A sensor node is composed of four basic components : a processing unit, a sensing unit, a power unit and a transceiver unit. A brief description of each component is provided below:

- The processing unit, usually associated with a small storage unit, is responsible for data processing and other assigned tasks by collaborating with other sensor nodes.
- The sensing unit is composed of two subunits: sensors and analog-to-digital converters (ADC). Sensors produce analog signals based on their observations. Those signals are converted into digital signals by ADC and are turned into the processing unit.
- The power unit is responsible for supplying powers from power generators such as battery or solar cells.
- The transceiver unit connects the node to the network.

In addition to these basic components, sensor nodes can be equipped with location finding systems and mobilizer depending on specific applications. Location finding systems (such as GPS) spots a location with a high accuracy. A mobilizer helps a sensor to move to a required position.

Some of the most important characteristics [19, 15, 14] of a sensor node are listed below:

- All of the components of a sensor node fit into a tiny module smaller than even a cubic centimeter.
- A node consumes extremely low power and operate in high volumetric densities.
- A sensor node is low cost, unattended and dispensable.
- A sensor node operates autonomously and is adaptive to the environment it is deployed.

### 2.3.2 Sensor Network Topology

Sensor nodes are deployed in a sensor fields in tens of thousands. Depending on the area of the sensor field, the maximum node density can be as high as 20 nodes/ $m^3$  [12]. There are three phases of sensor deployment and topology maintenance.

- Pre-deployment and deployment phase: Sensor nodes can be deployed in two possible ways. One way is to drop nodes in a mass from air plane, artillery shell, rocket or missile. The other way is to place nodes one by one in selective points with help of humans or robots.
- Post-deployment phase: After the deployment of sensors, the topology changes dynamically due to the change in position of sensors, reachability, malfunctioning, nature of the assigned task, available energy and node failures [19].
- Redeployment of additional nodes: Additional nodes can be at any time to replace malfunctioning or due to change in task dynamics.

### 2.3.3 Power Consumption

A sensor node is equipped with a very limited power source ( $\approx 0.5$  Ah, 1.2 V) [3]. Sensor nodes being unattended and unreplenished in most of the applications, the network lifetime is limited. Therefore, proper power management is the key to maximize network lifetime.

Power consumption of a sensor node occur due to three following events: sensing, data transmission and data processing. Data transmission consumes maximum power. Therefore, designing power-aware routing protocols and routing algorithms is one of the most significant challenges in sensor networks research.

### 2.3.4 Transmission Media

In most wireless sensor networks, radio is the main communication mode. The  $\mu AMPS$  wireless sensor node described in [12] uses a Bluetooth compatible transceiver of capacity 2.4 GHz with an integrated frequency synthesizer. Sensor nodes described in [13] uses a 916 MHz single channel transceiver. The Wireless Integrated Network Sensors (WINS) architecture [15] also uses radio links for communication. Unlike these systems, the autonomous system called Smart Dust mote [14] uses optical medium for transmission. This medium requires a line of sight between the receiver and the sender. Another possible mode is infrared communication which too requires a line of sight. Infrared based transceivers are cheaper, license free and robust to interference of electronic devices.

## 2.4 Protocol Stack

The protocol stack of a sensor network is composed of five basic layers e.g., *physical layer*, *data link layer*, *network layer*, *transport layer* and *application layer*. [3] The other components of a protocol stack are power management plane, mobility management plane and task management plane. Each layer and each plane perform different tasks of information collection and dissemination. Brief descriptions of each stack and plane are provided below.

### 2.4.1 Physical Layer

While frequency generation and signal detection is mainly connected to the underlying hardware and transceiver designs, the physical layer is directly responsible for power efficient signal propagation and modulation. We provide a brief discussion below:

- Energy aware propagation is an important factor in the design of the physical layer. Generally, the power required to propagate a signal over a distance  $d$  is proportional to  $d^n$ ,  $2 \leq n \leq 4$  [35, 36]. Although some research has been done, this area is still vastly unexplored.
- Reliable communication in a sensor network requires a good modulation scheme. In [12], binary and M-ary schemes have been compared. According to this study, binary

scheme is more energy efficient under startup power dominant conditions. [37] presents a low power direct-sequence spread specturn modem architecture for sensor networks.

## 2.4.2 Data Link Layer

The main component of the data link layer is medium access control (MAC). The data link layer is also responsible for performing error control, data frame detection, multiplexing of data stream etc. We discuss some of the important MAC layer protocols specially designed for wireless sensor networks as follows.

**Medium Access Control:** The general objective of a MAC protocol is to ensure a fair and effective sharing of the communication medium among sensor nodes while solving the problems of latency, deadlock and livelock. While there exist many MAC protocols for wireless ad hoc networks, those protocols do not meet the requirements of a sensor network due to the stringent resource constraints. For example, a cellular system is a infrastructure-based network with base-stations forming a wired backbone. MAC protocols for such systems have the primary objective of providing quality of service (QoS) and bandwidth efficiency. Base stations having unlimited power supply and the energy for mobile devices being replenishable, MAC protocols for such a system assume a dedicated resource assignment strategy. Clearly, MAC protocols for cellular systems are inadequate for energy constrained and distributed sensor networks which have no central controlling system as base stations.

Bluetooth is an infrastructureless short range (10 meter) wireless system with a star-shaped network architecture which uses time division multiple access (TDMA) [16]. The mobile ad hoc network (MANET) [17] is also an infrastructureless wireless system for which many MAC protocols have been proposed [20, 21, 22, 23, 24, 25]. But none of these MAC protocols meet the specifications of a sensor network due the following reasons:

- While size of a sensor network is usually much larger than bluetooth and MANET, the transmission range and radio range of a sensor network is much smaller than that of bluetooth and MANET.
- The primary objective of MAC protocols for both bluetooth and MANET is to ensure quality of service since both systems have adequate power supply.
- Topology changes are more frequent in a sensor network.

**Some Significant MAC protocols for Sensor Networks:** Existing MAC protocols for wireless sensor networks fall into two main categories: contention-based and based on reservation and scheduling. The standardized IEEE 802.11 [20], widely used for MANET, is not suitable for a sensor network due to its high energy consumption by nodes in idle mode. PAMAS [26]

is an improved version of standardized IEEE which aims at reducing energy consumption intelligently powering off nodes which are not actively transmitting or receiving packets. Sensor-MAC protocol [27] is further improvement over PAMAS which achieves further reduction in energy consumption and self-configuration. It makes node sleep periodically to reduce energy consumption due to idle listening. S-MAC, like PAMAS, also sets the radio to sleep during transmissions of other nodes. However, unlike PAMAS, it only uses in-channel signaling. S-MAC applies message passing to reduce contention latency for sensor-network applications that require store-and-forward processing as data move through the network.

A significant TDMA based MAC protocol for wireless sensor networks is proposed by Sohrabi and Pottie [53]. In this protocol, each node maintains a TDMA like frame, called super frame, in which a node schedules different time slots to communicate with its known neighbors, one per time slot. The protocol avoids interference between adjacent links by assigning different channels, i.e., frequency (FDMA) or spreading code (CDMA), to potentially interfering links. However, the actual multiple access is accomplished by FDMA or CDMA since the super frame, unlike TDMA, does not prevent two interfering nodes from accessing the medium at the same time. A drawback of the scheme is its low bandwidth utilization. Piconet [29] is an architecture designed for low-power ad hoc wireless networks which provides multiple access by letting a node broadcast its address before it starts listening. A node cannot communicate until it receives the neighbors broadcast. Piconet too puts idle nodes to sleep reduce energy consumption. Woo and Culler [30] proposed an adaptive rate control mechanism based on carrier sense multiple access (CSMA). Unlike S-MAC which trades of per node fairness for energy savings, this protocol aims at achieving fair bandwidth allocation to all nodes in a multi-hop network.

### 2.4.3 Transport Layer

This layer maintains the data flow according to application-specific requirements of sensor networks and provides access to the system through the internet or any other external network. The traditional TCP/IP or UDP protocol widely used in wired networks cannot be applied directly to wireless networks due to the following reasons:

- In ordinary IP networks, IP addresses are assigned either manually or using some dynamic mechanism which are impossible to implement in a sensor network due to network size and energy constraint.
- TCP/IP protocols incur large header overheads which is infeasible for resource constrained sensor nodes.

- While TCP/IP protocols use address centric routing, most applications of sensor networks require data-centric routing and addressing.
- The end-to-end acknowledgment and retransmission scheme employed by TCP causes expensive retransmissions along every hop of the path between the sender and the receiver, thereby making it infeasible for a sensor network.

Akyldiz et al. in [3] proposes an approach using TCP splitting in which communication between a sink and the end user will be performed using TCP/IP via internet or satellite while ordinary sensor nodes will communicate with the sink by pure UDP using data-centric addressing. Dunkels et al. in [31] proposes a number of modifications of the traditional TCP/IP protocol to make it viable for a wireless sensor network. They present mechanisms such as spatial IP address assignment, shared context header compression, application overlay routing, and distributed TCP caching (DTC).

#### 2.4.4 Network Layer

The network layer is responsible for routing the data supplied by the transport layer. Traditional ad hoc routing protocols can not be directly applied to a sensor network due to the following reasons:

- The number and density of sensor nodes in a network can be several orders of magnitude higher than those in a traditional ad hoc network.
- Sensor nodes are stringently constrained in power, computational capacities, memory and size. Moreover, these nodes are unattended and cannot be replenished, thereby limiting the network lifetime. Therefore, energy awareness plays a significant role in designing routing schemes for sensor networks. Ad hoc network routing protocols need not be energy aware.
- Sensor nodes are more prone to failures.
- Topology of a sensor network changes very frequently.
- Sensor nodes may not have global identifications due to large overhead.
- Routing in a sensor network is mainly data-centric and hence requires attribute-based addressing.

Energy efficient routes can be found based on residual energy in the nodes. In [3], some metrics such as maximum residual energy route, minimum energy route i.e., route that consumes minimum energy, minimum hop route, maximum minimum residual energy route have been proposed for finding energy aware routes. In this chapter, we discuss some of the basic routing techniques and schemes. In chapter 5, more recent works on routing will be presented.

**Flooding and Gossiping:** In this technique, each node broadcasts a received packet to all its neighbors unless the node is the destination of the packet. The main advantage of flooding is that it is easy to implement and does not require any expensive topology maintenance. However, there are several drawbacks as follows:

- **Overlap and Implosion:** Data overlap occurs when two or more nodes from the same region send the same information. Data overlap triggers data implosion by making two or more nodes sending the same information to a common node.
- Flooding technique is not energy aware.

Gossiping [33] is an improvement over the primitive flooding where a node sends information to a randomly selected neighbor, thereby reducing data redundancy resulted from implosion. However, a pure gossiping technique incurs long routes.

**Sensor Protocols for Information via Negotiation (SPIN):** This family of adaptive protocols [32] overcome the disadvantages of flooding technique by exploiting a basic idea: each node first sends a descriptor of the data it intends to send eventually. SPIN uses three types of messages: ADV, REQ and DATA. Before sending DATA, a node sends meta-data ADV which describes DATA. If a neighbor wants to receive data after getting ADV, it sends a REQ to the corresponding node. This node then sends DATA, thus eradicating the chance of any overlap or implosion.

**Sequential Assignment Routing:** Sequential assignment routing (SAR) [34] creates paths from sensors to the sink using the following two metrics:

- Energy resources is the estimated number of packets that can be sent by a sensor node via an exclusive path.
- The higher an additive QoS, the worse is QoS.

Each node selects its own path to the sink, thus creating multiple trees with one-hop neighbors of the sink as the roots. [34] describes some more algorithms i.e., SAMACS, EAR, SWR.

Directed Diffusion: In this paradigm for information dissemination [19], the sink floods the whole network with attribute-based queries called *interests*. An interest contains a timestamp field and several *gradients* fields. When an interest propagates from the sink to the sources, these gradients set up reverse paths from the sources to the sink. Each sensor node stores interests in its cache. As soon as a sensor collects data for an interest, it sends the data along the path set up by the gradients. Data aggregation can be performed at every sensor node.

Low Energy Adaptive Clustering Hierarchy (LEACH): LEACH [18] is a clustering-based protocol which minimizes energy consumption by dividing the whole network into clusters each with a corresponding cluster-head responsible for aggregating data and communicating with the sink. Thus, individual sensor nodes communicate only with their corresponding cluster heads, thereby reducing energy consumption. LEACH operates into two phases: the setup phase during which cluster heads are selected at random and the steady phase when ordinary sensors sense data and transmit to the cluster head. TDMA is used for medium access control in steady phase. The setup phase is usually shorter than the steady phase and both the phases are performed periodically.

## 2.5 Conclusion

This chapter presents a brief overview of wireless sensor networks with corresponding citations. The structure of this chapter is similar to that in [3]. In this chapter, we do not present a complete literature review for the problems considered for this thesis. Instead we provide comparative reviews of related literature in the corresponding chapters.

# Chapter 3

## Bootstrapping Problem and Key Pre-distribution

### 3.1 Introduction

Nodes in a sensor network are typically deployed in an ad-hoc manner into arbitrary topologies before self-organizing into a multihop network for collecting data from the environment and forwarding to the base station or sink [3],[40]. Establishing a secure communication infrastructure among a collection of arbitrarily deployed sensor nodes is an important and challenging security issue (known as the *bootstrapping problem* [38]). Due to severe computational and memory constraints, symmetric key cryptography is the most feasible encryption mechanism for node to node communication. However the high energy-cost of routing makes traditional methods of key exchange and key distribution protocols based on trusted third party mechanisms difficult to implement.

Since bootstrapping should not rely on pre-existing trust associations between fixed sensor nodes or the availability of an on-line service to establish these trust associations, an attractive alternative for secure encrypted communication between adjacent sensor nodes is key pre-distribution, i.e. pre-installing a limited number of keys in sensor nodes prior to actual deployment. Key pre-distribution is also challenging since ad-hoc network deployment makes it impossible to pre-determine the neighborhood of any node. Yet key distribution schemes must *ensure good network connectivity*(through key sharing) and *resilience to node/key capture by the enemy* even with limited number of keys per node. A trivial pre-distribution solution is to have a single secret key shared among all nodes. While this solution keeps the network fully connected (every node can communicate with every other node) and scalable (new nodes can be added without any keying overhead), it provides

extremely poor resiliency to enemy attack. At the opposite end of the spectrum, one can have each pair of nodes sharing a distinct key. This solution provides both high connectivity and high security but is very memory-intensive and not scalable.

There have been several recent works on key pre-distribution [41, 38, 47, 65, 44]. The pioneering paper in [41] proposes a simple, scalable probabilistic key pre-distribution scheme in which a certain number of keys are drawn at random from a (large) key pool and distributed to sensor nodes prior to their deployment. Note that due to the random distribution of keys and ad hoc deployment of sensors, there is a non-negligible probability of a disconnected logical graph. The degree of connectivity of the resultant sensor network under a given key pre-distribution scheme is therefore an important performance metric. There is also a strong *negative* correlation between network connectivity and security. Adversaries that capture nodes can gain complete information about the keys stored at the node in the worst case. Thus in order to make the network less vulnerable to node/key capture the overall key pool size must be large. Since individual sensor nodes have limited memory for key storage, this reduces the probability of having a large number of shared keys between neighboring sensors.

Good solutions for key pre-distribution must be memory-efficient and scalable, simultaneously ensuring that (a majority of) the network is connected through secure communication links and provide high resiliency to enemy attack so that the capture of a few sensor nodes does not (severely) compromise network communication. In this dissertation, we propose a novel solution to the key predistribution problem (labeled 2-Phase key predistribution) that exploits the connectivity and capture-resiliency properties of loading sensor nodes with a combination of *randomly derived* and *inherited* keys. We evaluate our solution by analytically developing novel quantitative metrics that measure the key predistribution schemes' security-performance tradeoffs in terms of the network resiliency to node/key capture, the number of available secure links and the key (memory) requirement per node for a given level of connectivity. We compare the network connectivity and security performance and show analytically and through simulations that the proposed 2-Phase scheme strongly favors highly secure large-composite key communication and is more resilient to node capture than the random scheme. We first show analytically that the invulnerability of an arbitrary  $q$ -composite communication link to any number of node captures is higher in our scheme. We also derive analytical results for measuring the vulnerability of a  $q$ -composite link to single-node capture assuming adversaries who can use captured-key knowledge network-wide as well as locally and show that the 2-Phase scheme is more resilient. Finally, we present simulation results that show the number of exclusive keys shared between two nodes is higher while the number of  $q$ -composite links compromised when a given number of nodes are captured by the enemy is smaller under the 2-Phase scheme.

The chapter is organized as follows: In section 2, security threats in a sensor network are briefly discussed. In section 3 and 4, we define and discuss the problem of bootstrapping

and its solutions respectively. Section 5 provides literature review on key predistribution. Section 6 describes the proposed 2-Phase scheme. Section 7 outlines our metrics for measuring security-performance tradeoffs. Section 8 contains some analytical connectivity results with Section 9 containing analytical results on security of communication edges under node capture. Section 10 describes simulation results followed by some implementation issues and Conclusions in Section 11.

## 3.2 Bootstrapping Problem in a Wireless Sensor Network

Establishing a secure communication infrastructure among a collection of arbitrarily deployed sensor nodes is known as the *bootstrapping problem* [38]. The main challenge for the bootstrapping problem is key agreement i.e., how to set up secret keys among communicating nodes. Along with initiating a secure infrastructure, a bootstrapping protocol must also allow additional nodes deployed at a later time to join the network securely. This is a challenging problem due to resource constraints of a sensor network.

### 3.2.1 Requirements and Challenges of a Bootstrapping Protocol in Sensor Networks

A solution for the bootstrapping problem in a sensor network must fulfill the following requirements:

- **Secure communication:** Communication between nodes must be properly encrypted. Links must be invulnerable to node capture i.e., capture of one or more nodes should not expose communication throughout the other links of the network to the enemy. Unauthorized nodes must not gain entry to the network and access to information disseminating through the existing links.
- **Resource efficient:** The computational expense and storage requirement must be low.
- **Scalability:** Additional nodes deployed at a later time must be able to establish secure communication with existing nodes.

However, due to hardware and software constraints of a sensor network, the bootstrapping problem in a sensor network becomes extremely complicated and challenging. We briefly discuss some of the major constraints here:

- **Lack of knowledge on network configuration:** Deployment of nodes by random scattering (e.g., from an airplane) makes it impossible to have a-priori knowledge of post-deployment network configuration. Even for a manually deployed sensor network, the huge network size makes predetermination of individual sensor locations extremely expensive. Therefore, security protocols must be developed assuming no prior knowledge of individual sensor positions.
- **Limited bandwidth and transmission power:** Typical sensor network platforms have limited bandwidth. For example, the UC Berkeley Mica platforms transmitter has a bandwidth of 10 Kbps [38]. Low transmission reliability [56] makes the communication of large blocks of data expensive.
- **Limited memory and processor capacity:** Storage and processor capacity are extremely low in a typical sensor node. For example, a SmartDust node has 8-bit, 4 MHz CPU, 512 RAM and 512 EEPROM [49].
- **High cost of providing tamper resistance mechanisms:** Hardware mechanisms to make a sensor node tamper-resistant is very high. Therefore, low-cost low-power sensor nodes deployed in hostile environments such as battlefields and public buildings are vulnerable to physical capture by an enemy. In the worst case, an enemy can get hold of all cryptographic keys, thereby making the whole communication system collapse.

### 3.3 Evaluation of Different Key Agreement Schemes with Respect to Sensor Networks

For general networks, there are three types of key agreement schemes: *Trusted Server Scheme*, *Self-enforcing Scheme* and *Key pre-distribution*. We provide brief descriptions of each scheme and their subsequent advantages and disadvantages.

#### 3.3.1 Trusted Server Scheme

This scheme assumes symmetric key cryptography for data encryption. Therefore, two nodes communicate using only one unique key shared by both of them. The trusted server method relies on a trusted base station to establish key agreement among pairs of nodes as described in [57, 58]. The steps of this protocol are as follows:

- A unique symmetric key is installed in the memory of each sensor node, prior to its deployment. The base station can access to all the node keys (either they are stored in its memory or the base station relays all communications to a secured workstation off site).
- After deployment, each node authenticates itself to the trusted secure base station using its symmetric key.
- After authentication, each pair of communicating nodes want to establish a shared secret session key with each other. Since a node pair do not share any secrets, they communicate with a trusted third party i.e., the base station.
- The base station generates a link key each for these node pairs and securely send the key to those nodes.

Advantages of this method are as follows:

- Memory efficient: Each node needs to store only one key for each of its neighbors and one for the base station.
- Perfect resiliency to node capture: Capture of any number of nodes does not expose communication in the rest of the network to the enemy.
- Simple Revocation of nodes: Having access to all the keys of all the nodes, the base station can securely transmit the revocation message to all the nodes that may be in communication with the revoked node. This revocation message is encrypted with the secret key that is shared only between the recipient node and the base station, and hence secrecy and authentication are ensured. To prevent any other nodes from establishing links with the revoked node, the base station simply needs to reject requests that involve the revoked node as a principal.

Disadvantages of this method are as follows:

- Highly energy expensive: Each node requiring to communicate with the base station several times, the total routing cost for key distribution and agreement is extremely energy expensive for a large sensor network where the base station can be very far away.
- Non-scalable: Addition of new nodes require several extra routing sessions to and from the base station, thereby further increasing energy cost.

- Insecure transport before key establishment: The base station and node pairs need to communicate before any authentication and key establishment takes place. This communication cannot be protected by any security mechanism.
- Compromise of base station: If an enemy compromises a base station, it can potentially get hold of all cryptographic information, thereby making the whole communication system collapse.

### 3.3.2 Self-enforcing Scheme

This scheme assumes public key cryptography for communication among nodes. The key steps of a key agreement process using public key cryptography for a sensor network [59] are as follows:

- First, a public/private key pair  $(K_M, K_M^{-1})$  is generated.
- Then for each node, its own public/private key pair is generated. This key pair, along with the master public key  $K_M$  and the master keys signature on the node's public key are installed in the node's memory prior to deployment.
- After deployment, nodes exchange their public keys and master key signatures. Verification of the master key signature using  $K_M$  authenticates the corresponding public key.
- Once public keys of two nodes are verified, a symmetric link key is generated and sent to both the nodes.

Advantages of this procedure are as follows:

- Perfectly resilient against node capture: Capture of any number of nodes cannot expose other links to the enemy.
- Fully scalable and memory efficient: Clearly deployment of additional nodes does not increase any extra routing overheads. Furthermore, each node has to store only a very small number of keys.
- Revocation of compromised nodes: A central authority can broadcast a revocation message signed by the master key. Nodes receiving this message can easily authenticate the message by using the master public key and ignore future messages from the revoked nodes. If the master key needs to be changed, a new master key signed by the old master key can be unicast to legitimate nodes. The keys of revoked nodes cannot be authenticated using this new master key.

However, all known public key cryptography mechanisms such as Diffie-Hellman key agreement [60] or RSA [61] involve intensive computations of complex mathematical functions. Sensor nodes having very low-power low-speed CPU, implementation of these mechanisms in a sensor network is practically impossible.

### 3.3.3 Key Pre-distribution

This scheme is based on symmetric key cryptography. A limited number of keys are pre-installed in each node prior to their deployment. The main steps of the key agreement scheme [41] using key pre-distribution are as follows:

- Nodes are deployed after keys are pre-installed in their memory. Each key has a key-identifier.
- After deployment, a *shared-key discovery phase* starts. During this phase, each node broadcasts the list of its key-identifiers. Note that this broadcast does not provide an enemy with any extra opportunity of attack. An enemy has to physically capture a node in order to get hold of actual keys. Two nodes which are within the communication ranges of each other establishes a link if they share a common key. Thus the topology established during this phase may be different from the original topology established prior to shared key discovery.
- During the *path establishment phase*, a path-key is assigned to a pair of nodes which are in the communication range of each other but are connected by two or more links. A path key may not be generated by the sensor nodes. Each node can have unused keys in the key ring which can be used as a path key.

Note that depending upon the process of pre-distributing keys, shared-key discovery and path establishment may not be needed to take place. For example, these phases are not required if a single key is used network-wide.

This scheme has the following advantages:

- Ease of implementation: This scheme is the most suitable method for a sensor network since it does not involve any extra computational or routing cost.
- Memory efficiency and scalability: This scheme can be made very memory efficient by limiting the number of keys that are to be pre-installed. This scheme is clearly scalable since it does not incur any additional cost for performing key agreement with additional newly deployed nodes.

- **Simple Revocation:** Revocation of a compromised node can be done by broadcasting a message to all the other nodes. The message should contain the list of all key identifiers possessed by the revoked key.
- **Resiliency to node capture:** Resiliency to node capture varies depending upon the process by which the keys are generated and distributed. There are several key pre-distribution schemes in the literature each providing different degree of resiliency. While it is not possible to provide 100 percent resiliency, the probability of vulnerability due to node capture can be minimized.

The only disadvantage of this scheme is that the post-deployment logical neighborhood of a node cannot be determined deterministically. Therefore, logical topology established after the shared-key discovery phase might be disconnected. However, the probability of having a connected graph after shared-key discovery can be highly improved if a key pre-distribution scheme increases the probability of two nodes having at least one common key.

Although key pre-distribution technique is quite attractive for key agreement in a sensor network for its highly resource (memory, energy and processor capacity) efficient nature, there is a performance trade-off among resiliency to enemy attack, connectivity and resource efficiency. The following two extreme cases illustrate the performance tradeoff.

**A Single Network-wide Key:** The simplest method of key pre-distribution is to make all the nodes share a single unique key. The advantages of a single key are the following:

- Memory requirement is minimum.
- Connectivity is guaranteed.
- Key discovery/key exchange is not required, thereby saving node energy.

However, this scheme provides no resiliency to enemy attack. Capture of a single key will expose the single key to the enemy, thereby making the whole communication system collapse. Therefore, in spite of its resource efficiency, this scheme is infeasible for a sensor network where low-cost nodes do not have tamper resistant hardware mechanisms.

**Pairwise Single Keys:** In this approach, every node shares a unique symmetric key with every other node in the network. Every node thus needs to store  $(N - 1)$  keys,  $N$  being the network size. After deployment, a node must verify the identity of the node that it wants to communicate with. This can be accomplished with a challenge/response protocol. The major

advantage of this method is that it provides perfect resiliency to node capture since each communication link uses a unique key to encrypt the data. However, each node requiring to store  $N - 1$  keys, this method is extremely memory inefficient and non-scalable. Therefore, this too is not feasible for a large sensor network.

## 3.4 Literature Review

There have been several recent works on key pre-distribution [41, 38, 47, 43, 44, 63, 64, 65, 66]. The pioneering paper in [41] proposes a simple probabilistic key pre-distribution scheme. [38, 47, 45] propose key management protocols where keys can be issued multiple times out of the key pool. [63, 64, 65, 66], on the other hand, propose different methods for *pairwise* key distribution assigns a unique key to each pair of nodes. In this section, we provide brief overviews of all these protocols.

### 3.4.1 Random Key Pre-distribution

This key pre-distribution scheme [41] has the following steps:

- A pool of  $L$  keys with key identifiers is generated during the *initialization phase*. The number of keys in the key pool is chosen such that two random subsets of size  $k$  selected from the pool will share at least one key with some probability  $p$ .
- For each node,  $k \ll L$  keys are drawn at random from the pool and are installed into the memory of the node. The set of the keys in a node is called its *key ring*.
- The shared-key discovery and the path establishment phase take place consecutively as described above. A (logical) graph with secured communication links is finally formed.
- Due to random distribution of keys and ad-hoc deployment of sensors, there exists a chance that the logical graph is disconnected, in which case sensor nodes perform *range extension* [38] until a connected logical graph is created. However, in terms of node energy, this procedure is quite expensive for a sensornet.

Characteristics of this procedure are stated below:

- The main advantage of this scheme is that it is very easy to implement and flexible. By choosing proper key pool size and key ring size, desired level of memory efficiency and connectivity can be achieved. However, 100 percent connectivity cannot be guaranteed. If the network detects that it is disconnected, sensor nodes may perform range extension

by increasing their transmission power, or sending a request to their neighbors to forward their communications for a certain number of hops until a connected graph is formed. A useful way for a node to detect if a network is connected is by checking if it can perform multi-hop communication with all base stations. However, connectivity detection and range extension can be very expensive for a resource constrained sensor network.

- The broadcast-based key discovery methods have the disadvantage that a casual eavesdropper can identify the key sets of all the nodes in a network and thus pick an optimal set of nodes to compromise in order to discover a large subset of the key pool. A more secure, but slower method of key discovery could utilize client puzzles such as a Merkle puzzle [62]. Each node could issue  $k$  client puzzles (one for each of the  $k$  keys) to each neighboring node. Any node that responds with the correct answer to the client puzzle is thus identified as knowing the associated key.
- The probability of connectivity can be increased by selecting right parameters. However, this paper does not provide any analytical framework to compute resiliency to enemy attack or to analyze the tradeoff between connectivity and security.

### 3.4.2 $q$ -Composite Communication Scheme Based on Random Key Pre-distribution

In [38], the authors have presented new communication scheme using the random key pre-distribution scheme of [41] as a basis. Their  $q$ -composite scheme requires that two adjacent communicating nodes must have at least  $q$  keys in common. The key steps are summarized as follows:

- During the initialization phase prior to deployment, the deployer/central authority picks a set  $L$  of random keys (key pool) out of the total key space.  $k$  keys are selected at random from the key pool and are installed in the memory of each node.
- During the shared-key discovery phase, each node identifies the neighbor nodes with which it shares at least  $q$  keys. A new communication link key  $K$  is then generated as the hash of all shared keys. The keys may be hashed in some canonical order, for example, based on the order they occur in the original key pool. This can be done by giving all keys in the key pool numerical identifiers sorted in some manner. However, this induces a little computational overhead.

The authors study that there is a tradeoff between the connectivity and the security which depends on the parameters such as the key pool size, the key ring size and the probability of connectivity. For example, a small key pool size increases connectivity but also induces a high number of common keys shared among a number of nodes, thereby increasing link vulnerability. On the other hand, a key pool size too large is highly probable to create a disconnected graph, thereby disrupting communication. Therefore, one way to solve this problem is to fix the value of one parameter and optimize the value of the other. In this paper, the authors assume that they are provided with a threshold value  $p$  such that the probability of any two nodes sharing at least  $q$  keys, ( $q$  pre-specified) is at least  $p$ . Note that this probability is an index of connectivity. Given this index, the authors maximize the key pool size. Formally, given key ring size  $k$ , minimum key overlap  $q$ , and minimum connection probability  $p$ , the authors select the largest key pool size such that  $p_{connect} \geq p$ , where  $p_{connect}$  is the probability that two nodes share at least  $q$  keys. This paper experimentally studies resiliency to enemy attack provided by the  $q$ -composite scheme and shows that the resiliency increases as the value of  $q$  decreases. However, the paper does not provide any analytical framework to compute this resiliency. The authors provide a multipath key reinforcement scheme to strengthen the security of an established link key by establishing the link key through multiple paths. This method applied in conjunction with the basic random key scheme yields huge resiliency against node capture by trading off some network communication overhead.

### 3.4.3 Evaluation of Random Key Pre-distribution Using Deployment Knowledge

In [47], the authors evaluate the random key predistribution scheme under a variety of non-random node deployment probability distributions. However, for many realistic sensor networks as visualized in Smart Dust [51], node deployment into known topologies is an unrealistic assumption.

### 3.4.4 Sub-vector Based Deterministic Key Pre-distribution Scheme

For sensor networks with significant memory constraints, [45] proposes a *deterministic* subvector based key pre-distribution approach based on distributed agreement using vector spaces and quorums [48]. The scheme has the deterministic property that any two nodes sharing a key under a given mapping (out of several possible mappings) from sensor nodes to keys share exactly two keys (that are unique to these two nodes) in this mapping. For a given network deployment, two physically adjacent sensors can encrypt messages using shared keys under one or more of these mappings, where each mapping yields a 2-composite key. It can be

shown that any ad hoc deployment of sensor nodes yields very high connectivity with low key storage requirements per node.

### 3.4.5 Random Pairwise Key Scheme

This scheme, proposed by Chan, Perrig and Song in [38], is summarized as follows:

- During the initialization phase prior to deployment, a total  $\frac{k}{p}$  unique node identities are generated, where  $k$  is the key ring size and  $p$  is the probability that two nodes can communicate securely. For each node in the network,  $k$  other distinct node IDs are selected at random and a pairwise key is generated for each pair of nodes. The key is stored in both nodes key rings, along with the ID of the other node that also knows the key. Note that  $k$  and  $p$  must be so chosen that the value of  $\frac{k}{p}$  is at least equal to the network size.
- After deployment, a key-setup phase takes place. During this phase, each node broadcasts its ID to its neighbors. A neighbor communicates back if it shares a common pairwise key for communication. These two nodes then mutually verify their key using a cryptographic handshake protocol.

### 3.4.6 Single-Space Pairwise Key Schemes

Schemes presented in [63, 64] propose that a sensor node  $i$  must store unique public information  $U_i$  and private information  $V_i$ . During the key setup phase, neighboring nodes exchange public information, and node  $i$  computes its key with node  $j$  with  $f(V_i, U_j)$ , where  $f$  is a polynomial such that  $f(V_i, U_j) = f(V_j, U_i)$ . Both schemes are  $\lambda$ -secure i.e., if the size of the set of compromised sensor nodes is less than  $\lambda$ , all other links in the network are invulnerable.

### 3.4.7 Multi-Space Pairwise Key Schemes

Multi-space pairwise key schemes proposed in [65, 66] enhances the security of a single-space scheme by combining it with the basic random key pool scheme. The setup server randomly generates a pool of  $L$  key spaces each of which has unique private information. Each sensor node will be assigned  $k$  out of these key spaces. If two neighboring nodes have one or more key spaces in common, they can compute their pairwise secret key using the corresponding single space scheme.

### 3.4.8 Polynomial Based Communication Scheme

[65] presents an alternative model for secure sensor communication using polynomials rather than keys but the computational constraint of using such polynomials is not extensively evaluated.

## 3.5 Two-Phase Key Pre-distribution Mechanism

We now describe a novel key pre-distribution scheme (labelled 2-Phase) in which sensor nodes are preloaded with a combination of *randomly derived* and *inherited* keys. Our proposed 2-Phase scheme is motivated by the following key observations:

- From the connectivity point of view, the probability of having a common key between two nodes decreases as the key pool size increases under the random pre-distribution scheme. We observe that the (probabilistic) connectivity of the logical graph can be increased if we can ensure that each node *deterministically* shares some of its keys with some nodes (as in the subvector scheme [45]).
- We hypothesize that it is better from the security point of view to pre-distribute keys in a less-random fashion such that whenever a node shares a key with another node, it should be likely to share a larger number of keys with this node. If so, the resulting network should consist of high-composite links. Note that  $q$ -composite schemes are more secure with increasing  $q$ . If the adversary has obtained  $X$  keys (through the capture of one or more sensor nodes), the probability of determining the exact  $q$ -subset of  $X$  that is used by a given communicating sensor pair decreases exponentially with increasing  $q$ .

We now describe the key steps in the proposed 2-Phase key predistribution mechanism. Order the sensor nodes apriori in a logical queue and distribute keys in increasing order according to the rules below.

- The first node is assigned  $k$  keys drawn randomly from the key pool of size  $L$ .
- For every succeeding sensor node  $i$ ,  $k$  keys are distributed in two consecutive phases. First, node  $i$  receives a predetermined fraction  $f$  ( $1/k \leq f < 1$ ) of its  $k$  keys drawn randomly from the key space of node  $i - 1$ . The remaining  $(1 - f)$  fraction of  $k$  keys are then drawn randomly from the key pool of size  $L - k$ , *after excluding all  $k$  keys of node  $i - 1$  from  $L$ .*

The 2-Phase scheme is designed to be biased in favor of nodes sharing several keys with their immediate predecessors and successors, through direct inheritance as well as a random component. Intuitively, this key predistribution methodology should offer better secure connectivity in the logical graph by inducing the sharing of larger number of keys between nodes, thereby enabling  $q$ -composite communication for larger values of  $q$ . More suprisingly however, as we show in the security analysis section, this methodology also provides enhanced security under node capture/eavesdropping by allowing for more 'exclusive' key sharing between communicating nodes. The fraction  $f$  (called *inheritance ratio*) plays a significant part in the connectivity/security of the logical graph created after node deployment. Note that the random key predistribution scheme **is not** a special case of the 2-Phase scheme with  $f = 0$ , since we eliminate all  $k$  keys of the previous node from regardless of the value of  $f$ . We will shortly derive relationships between 'good' values of the various parameters  $k, L, f$  etc. Finally, the proposed 2-Phase scheme is scalable since new sensor nodes can be assigned keys according to this rule at any time.

Note that there is an implicit ordering of sensors based on their position in the logical queue which determines each nodes key set. Thus each node has a logical identifier which we will refer to as its LID. Storing a node's LID in memory is an implementation decision as there is an associated security-performance tradeoff. If LID's are stored, nodes can be restricted to forming communication links only with adjacent nodes whose LIDs are greater than a specified minimum and within a specified maximum LID distance. As shown later, this will encourage the formation of high-composite encrypted communication that are also less vulnerable to compromization in the case of node capture. Conversely, storing LIDs will enable the adversary to target nodes with specific LIDs (although their positions will still be unknown). Therefore this becomes an implementation issue.

### 3.6 Metrics for Measuring Security-Performance Tradeoffs

Since security mechanisms directly impact system performance, there is a strong need to develop a rigorous analytical framework for measuring the security-performance tradeoffs of arbitrary key distribution schemes. These tradeoffs can be represented as functions of individual metrics which measure the networks 'secure' connectivity in terms of the number of available secure links or paths, the memory requirement in terms of keys per node for a given level of connectivity and measuring resiliency of the network to node/key capture. In

this paper, we obtain some new analytical results on the security-performance tradeoffs of key predistribution schemes using the quantitative metrics outlined below. Results for the proposed 2-Phase scheme are compared with random key predistribution.

- Connectivity Metrics

- Logical sensor degree: We measure the logical degree of a node as the number of adjacent sensor nodes (in the logical graph) with which it shares at least one key. The higher the expected node degree, the better the connectivity of the logical graph. A high expected degree also implies a larger expected number of disjoint paths from any source to any destination. Multiple disjoint paths can be used to split communication and carry disjoint messages, thereby increasing overall data security. We show that nodes under the proposed 2-Phase scheme have higher expected degrees as compared to random key predistribution.
- Number of keys shared between any two neighboring nodes: This metric can be used to evaluate connectivity under  $q$ -composite key communication. We show that any two sensor nodes are expected to share more keys and are more likely to share  $q$  keys for any value of  $q$  (thereby enabling  $q$ -composite communication), as compared to random key predistribution.

- Security Metrics

- Exclusive Key Sharing: If two communicating nodes share one or more keys exclusively, then their communication is *invulnerable* to any number of node captures. Note that the exclusivity metric can be computed network-wide or with respect to a local cluster<sup>1</sup>. Network wide exclusivity between communicating nodes implies resilience against a powerful adversary who can capture nodes and use the captured key information *anywhere* in the sensor network. Alternatively, we can consider a weaker adversary who can use the key information only within the cluster of the captured node.
- Node Capture: We measure the impact of node capture on network security by considering the number of communication links that are no longer secure (i.e. only use keys from the captured key pool). We analytically determine bounds on the inheritance ratio  $f$  for which the 2-Phase scheme shows good resilience to network-wide as well as localized single-node capture and present simulation

---

<sup>1</sup>Typical sensor networks are organized into hierarchical clusters with cluster heads, such that each node is within wireless range of other nodes in the cluster [18]. Thus a compromised node can potentially eavesdrop on all intra-cluster communication.

results that show good network resilience to multiple-node capture as well. The expected number of links compromised in these cases is shown to be lower for the 2-Phase scheme as compared to the random scheme.

### 3.7 Secure Network Connectivity: Analytical Results

**Proposition 1** *Let  $l$  and  $i > l$  be any two nodes in the sensor network. The expected number of keys shared by  $l$  and  $i$  under the 2-Phase and Random schemes, respectively, are*

$$\begin{aligned} E_{l,i}^{2P} &= k \left( \frac{k}{L} + \left( \frac{fL - k}{L - k} \right)^{i-l} \left( 1 - \frac{k}{L} \right) \right) \\ E_{l,i}^{Rand} &= \frac{k^2}{L} \end{aligned}$$

**Proof:** The number of common keys between any two nodes under the random scheme is the standard hypergeometric distribution with parameters  $k$  and  $L$ , whose mean is  $k^2/L$ . For the 2-Phase scheme, let  $X_r$  be the number of keys in common between nodes  $l$  and  $l+r$ . Then we have,

$$\begin{aligned} X_{r+1} &= fX_r + (k - X_r) \frac{k - fk}{L - k} \\ &= X_r \frac{fL - k}{L - k} + k \frac{k - fk}{L - k} \end{aligned}$$

since after selecting an expected  $fX_r$  keys from the previous node, there are  $k - X_r$  keys of node  $l$  left in the random keypool of the current node.  $E_{l,i}^{2P} = X_{i-l}$  is the solution to the above recurrence relation with initial condition  $X_0 = k$ .  $E_{l,i}^{2P} > E_{l,i}^{Rand}$  as expected. ■

Thus to ensure  $q$ -composite connectivity between arbitrary nodes, a good choice is to select  $k$  and  $L$  such that  $q = k^2/L$ . Further, if  $f = k/L$ , then the expected number of common keys between any two nodes is identical under both schemes.

**Corollary 1** *The probability that any two nodes share at least  $q$  keys and the expected  $q$ -composite degree of a sensor node (i.e., number of neighbors with which it shares more than  $q$  keys) is higher under the 2-Phase key distribution scheme,  $f \geq k/L$ .*

As nodes are more likely to share multiple keys under 2-Phase, the probability of uncovering all such common keys (which is necessary to decipher data transmissions between the two nodes) can be shown to be lower and hence two-phase is more secure in this respect.

## 3.8 Network Resiliency against Enemy Attack: Analytical Results

In this section, we propose some quantitative metrics for measuring the security of communication links under enemy attack and analytically evaluate these metrics under different adversarial models. We assume an adversary that is able to capture nodes and obtain full knowledge of the captured node’s key space. We evaluate link security under a ‘network-wide’ adversary who can use knowledge of captured keys to compromise communication in any part of the network (regardless of the physical location of the captured node). Our results can be easily extended to analyze link vulnerability in the presence of a localized adversary who utilizes captured key knowledge locally, i.e. can compromise communication within a small neighborhood of the captured node (for example, its cluster as in LEACH [18]).

### 3.8.1 Vulnerability Under Multiple Node Capture: Key Exclusivity

We first evaluate the vulnerability of logical communication links in the sensor network to multiple node capture. An obvious metric for measuring this vulnerability is the degree of exclusivity of the keys used by any two neighboring nodes for setting up a communication link. Therefore we evaluate the probability of any two neighboring nodes containing exactly one *network-wide* exclusive key, the presence of which will render their communication link invulnerable to any number of (other) node captures <sup>2</sup>.

**Proposition 2 Key Exclusivity:** *In an  $N$  node sensor network, the probability that a given communication link between two arbitrary neighboring sensors is invulnerable to any number of network-wide node captures is given by:*

$$\begin{aligned}
 \text{2-Phase:} & \quad \frac{(1-f)^4}{1-(k/L)} \left(\frac{k}{L}\right)^2 \left(1 - \frac{k}{L} \left(\frac{1-f}{1-(k/L)}\right)\right)^{N-5} \\
 \text{Random:} & \quad \left(\frac{k}{L}\right)^2 \left(1 - \frac{k}{L}\right)^{N-2}
 \end{aligned}$$

*Link invulnerability is higher under the 2-Phase scheme for  $\frac{1}{k} \leq f \leq \frac{k}{L}$ .*

---

<sup>2</sup>In general, we can compute the probability of two nodes containing at least one exclusive key, but for all practical purposes this probability drops off extremely rapidly for more than one exclusive key. Hence we obtain a simple lower bound on invulnerability by focusing on the presence of a single exclusive key.

**Proof:** Let  $IV^{rand}$  and  $IV^{2P}$  denote the probability that two arbitrary neighboring sensor nodes  $i$  and  $j$  communicate using an exclusive key under the two key predistribution schemes. In the case of the 2-Phase scheme,  $i$  and  $j$  represent the LIDs of the communicating nodes. Consider a specific key  $a$  from the key pool.

For the random scheme, the probability that both nodes  $i$  and  $j$  possess key  $a$  is  $(k/L)^2$  while the probability that an arbitrary node  $l \neq \{i, j\}$  does not possess key  $a$  is  $1 - ((L-1)/\binom{L}{k}) = 1 - k/L$ . Hence the invulnerability of the link between nodes  $i$  and  $j$  under any number of node captures is given by:

$$IV^{Rand} = \left(\frac{k}{L}\right)^2 \left(1 - \frac{k}{L}\right)^{N-2} \quad (3.1)$$

For the 2-Phase scheme, the probability that key  $a$  is exclusive to nodes  $i$  and  $j$  is the probability that node 1 does not select key  $a$ , followed by all nodes up to node  $i - 1$  not selecting key  $a$  conditioned on the fact that their predecessor node did not select key  $a$ . Node  $i$  then selects key  $a$  given that node  $i - 1$  did not select it. Similarly all nodes after  $i$  conditionally do not select key  $a$  except node  $j$ .

Let  $P(1^c)$  denote the probability that node 1 does not contain key  $a$ ,  $P(1^c) = 1 - k/L$  since node 1 selects keys from the keypool first. Similarly, let  $P(l^c | (l-1))$  denote the probability that a node  $l$  does not contain key  $a$  given node  $l - 1$  contains it,  $P(l^c | (l-1)) = 1 - f$ , by definition. Finally, we have

$$\begin{aligned} P(l^c | (l-1)^c) &= \frac{\binom{L-k-1}{k-fk}}{\binom{L-k}{k-fk}} \\ &= 1 - \frac{k}{L} \left( \frac{1-f}{1-(k/L)} \right) \\ P(l | (l-1)^c) &= \frac{k}{L} \left( \frac{1-f}{1-(k/L)} \right) \end{aligned}$$

We now consider two cases (WLOG assume  $j > i$ ):

**Case 1:**  $j > i + 1$ :

$$\begin{aligned} IV^{2P} &= P(1^c)P(2^c | 1^c) \cdot \dots \cdot P((i-1)^c | (i-2)^c)P(i | (i-1)^c)P((i+1)^c | i) \\ &\quad P((i+1)^c | i) \cdot \dots \cdot P(j | (j-1)^c)P((j+1)^c | j) \cdot \dots \cdot P(N^c | (N-1)^c) \\ &= \frac{(1-f)^4}{1-\frac{k}{L}} \left(\frac{k}{L}\right)^2 \left(1 - \frac{k}{L} \left(\frac{1-f}{1-(k/L)}\right)\right)^{N-5} \end{aligned} \quad (3.2)$$

**Case 2:**  $j = i + 1$ :

$$IV^{2P} = (1 - f)^2 f \frac{k}{L} \left( 1 - \frac{k}{L} \left( \frac{1 - f}{1 - (k/L)} \right) \right)^{N-4} \quad (3.3)$$

Comparing Equations 3.2 and 3.1, the probability of two nodes having a network wide exclusive key (i.e. link invulnerability) is higher under the two-phase scheme as compared to the random scheme for  $\frac{1}{k} \leq f \leq \frac{k}{L}$ . ■

We can consider an alternative version of the 2-Phase scheme that provides much greater key exclusivity. The first step of key selection is the same as before, i.e. node  $i$  selects  $fk$  keys from the key space of node  $i - 1$ . However, in the second step, only the  $fk$  keys selected from node  $i - 1$  are excluded from keypool  $L$  before node  $i$  selects its remaining  $k - fk$  keys. For this modified 2-Phase scheme called 2PWR (2-Phase with replacement), we can show the following:

**Proposition 3 Scalable Comparative Exclusivity:** *The invulnerability of a communication edge under any number of node captures when keys are distributed using the 2PWR scheme is*

$$IV^{2PWR} = IV^{Rand} \frac{(1 - f)^4}{\left(1 - f \frac{k}{L}\right)^{N-1}}$$

*Thus link invulnerability under 2PWR outperforms the random scheme as the size of the sensornet  $N$  scales upward. This link invulnerability is maximized when*

$$f = \frac{(N - 1) \frac{k}{L} - 4}{(N - 5) \frac{k}{L}}$$

**Proof:** Using the same technique as in proposition 2 the probability of a given communication link  $(i, j)$  containing a network-wide exclusive key under the 2PWR scheme is given by:

$$\begin{aligned} IV^{2PWR} &= \frac{(1 - f)^4}{\left(1 - f \frac{k}{L}\right)^{N-1}} \left(\frac{k}{L}\right)^2 \left(1 - f \frac{k}{L}\right)^{N-2} \\ &= \frac{(1 - f)^4}{\left(1 - f \frac{k}{L}\right)^{N-1}} IV^{Rand} \end{aligned} \quad (3.4)$$

The value of  $f$  that maximizes the above term can then be found using elementary calculus. ■

While key exclusivity and (average network connectivity) are superior under the 2PWR scheme, the vulnerability of a link to single node capture is lower under the standard 2-Phase scheme as shown in the next section and hence we focus on that scheme for the rest of the paper. The choice of particular key distribution scheme with its associated security performance tradeoffs then becomes an implementation issue.

The next section contains some analytical results on edge vulnerability under localized node capture when the average node density of the sensor net (i.e size of a network cluster) is  $M$ . Simulation results on the number of exclusive keys per communicating node pair in a cluster are presented in Section 7.

### 3.8.2 Link Vulnerability Under Single Node Capture

We now consider the vulnerability of communication links in the sensor net to the capture of a single node by the adversary. We assume that the adversary does not possess any extra knowledge about the network topology and thus the capture of any given node by the adversary is equally likely.

Let  $i$  and  $j$  be any two communicating sensors in radio range and suppose the adversary captures node  $l$ . The vulnerability of edge  $(i, j)$  is the expectation (over all network nodes) that node  $l$  contains all the keys in common between  $i$  and  $j$ . We can thus define a network-wide vulnerability metric  $VC$  for arbitrary edges in the sensor net as follows:

$$VC = \sum_{l \neq i, j} P[\text{node } l \text{ is captured}] \cdot P[l \text{ contains all keys used to communicate over } (i, j)] \quad (3.5)$$

Before describing our vulnerability results, we first prove the following useful lemmas.

**Lemma 1** *Let  $i, i - x, i + x$  be arbitrary nodes in a sensor net in which keys have been predistributed according to the 2-Phase scheme. Let  $Z$  be any subset of keys from the keyset of node  $i$ ,  $|Z| \leq k$ . The probability distribution of the number of keys from  $Z$  that appear in nodes  $i - x$  and  $i + x$  are identical and dependent only on the LID difference  $x$  for both 2PWR as well as 2POR.*

**Proof:** Clearly, the total number of common keys between nodes  $i - 1$  and  $i$  and between nodes  $i$  and  $i + 1$  follows the same probability distribution, since they are obtained in an identical manner through inheritance followed by keys from the random pool. Thus the number of common keys from any subset  $Z$  of  $i$ 's keys also follows the same distribution in  $i - 1$  and  $i + 1$ . The lemma follows by induction on  $x$ . ■

The following statements follow directly from lemma 1 since the number of keys in common between any two nodes under 2-Phase depends only on their LID difference.

**Corollary 2** *Let  $i$  and  $j$  be two arbitrary nodes in the sensornet in which keys are pre-distributed according to the 2-Phase scheme,  $j > i$ . Consider nodes  $i - t$  and  $j + t$ ,  $t \geq 1$ .*

- *The number of keys in common between nodes  $i - t, i$  and  $j, j + t$  follows identical probability distributions.*
- *Suppose nodes  $i - t, i$  ( $j, j + t$ , resp.), share exactly  $\beta$  keys,  $0 \leq \beta \leq k$ . Then the number of keys from the remaining keyset of  $i$  ( $j$ , resp.) present in node  $j$  ( $i$ , resp.) follow identical probability distributions.*

*The above statements also hold for nodes  $i, i + y$  and  $j - y, j$ , where  $y \leq \lceil (j - i)/2 \rceil$ .*

**Lemma 2** *Let  $l, i$  and  $j > i$  be any three nodes in a sensornet, such that  $i$  and  $l$  share exactly  $\beta$  keys,  $0 \leq \beta < k$  and  $0 \leq l \leq i + \lceil (j - i)/2 \rceil$ . Let  $Z$  denote the set of remaining keys in node  $i$ ,  $|Z| = k - \beta$ .  $E_Z$ , the expected number of keys from  $Z$  that are present in  $j$  is given by*

$$E_Z = \begin{cases} (k - \beta) \left( \frac{k}{L} + \left( \frac{fL - k}{L - k} \right)^{j-i} \left( 1 - \frac{k}{L} \right) \right) & \text{if } l < i \\ (k - \beta) \frac{k}{L} \left( 1 - \left( \frac{fL - k}{L - k} \right)^{j-l} \right) & \text{if } i < l \leq i + \lceil (j - i)/2 \rceil \end{cases}$$

**Proof:** Let  $X_r$  represent the expected number of keys from keyset  $Z$  in node  $i + r$  (if  $l < i$ ) and  $l + r$  (if  $i < l < j$ ).  $E_Z$  is obtained by solving the recurrence relation

$$\begin{aligned} X_r &= fX_{r-1} + (k - \beta - X_{r-1}) \frac{k - fk}{L - k} \\ &= X_{r-1} \left( \frac{fL - k}{L - k} \right) + (k - \beta) \frac{k - fk}{L - k}. \end{aligned}$$

with initial condition  $X_0 = k - \beta$ , if  $l < i$  and  $X_0 = 0$ , if  $i < l \leq i + \lceil (j - i)/2 \rceil$ . ■

Assume that node  $l$  is captured and let  $CR_l$  be the Bernoulli random variable indicating whether capture of  $l$  reveals all common keys between communicating nodes  $i$  and  $j$ . Denote  $PCR_l : Pr.[CR_l = 1]$ . We now state our first proposition on sensornet vulnerability under single node capture.

**Proposition 4** *The probability of a given communication link between neighboring sensors  $i$  and  $j$  being compromised by the capture of an arbitrary node  $l \neq i, j$  is given by*

$$\begin{aligned}
PCR_{i-t}^{2P} = PCR_{j+t}^{2P} &\leq P_{i-t,i,k}^{2P} + \sum_{\beta=0}^k P_{i-t,i,\beta}^{2P} \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \left(1 - f\left(\frac{k}{L} + B^{j-i}\left(1 - \frac{k}{L}\right)\right)\right)^{k-\beta} \\
&\quad \text{for } t \geq 1 \\
PCR_{i+t}^{2P} = PCR_{j-t}^{2P} &\leq P_{i,i+t,k}^{2P} + \sum_{\beta=0}^k P_{i,i+t,\beta}^{2P} \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \left(1 - f\left(\frac{k}{L}(1 - B^{j-i-t})\right)\right)^{k-\beta} \\
&\quad \text{for } 1 \leq t < \lceil \frac{j-i}{2} \rceil \\
PCR_l^{Rand} &= P_{l,i,k}^{Rand} + \sum_{\beta=0}^k P_{l,i,\beta}^{Rand} \frac{\binom{L-k+\beta}{k}}{\binom{L}{k}} \quad \forall l \neq \{i, j\}.
\end{aligned}$$

where  $P_{l,i,\beta}^{2P}$  and  $P_{l,i,\beta}^{Rand}$  denote the probability that nodes  $l$  and  $i$  share exactly  $\beta$  keys under the specified key distribution scheme and  $B = \frac{fL-k}{L-k}$ .

**Proof:** Without loss of generality, assume  $1 \leq l \leq i + \lceil \frac{j-i}{2} \rceil$  and let nodes  $l$  and  $i$  share exactly  $\beta$  keys. Let  $Z$  denote the set of remaining keys in node  $i$ ,  $|Z| = k - \beta$ . Under the 2-Phase scheme, node  $j$  can first obtain keys from keyset  $Z$  through inheritance from its predecessor, node  $j - 1$ , and then from the random keypool of size  $L - k$  (obtained after removing the  $k$  keys of node  $j - 1$ ). Let  $jmZ$  be a random variable denoting the number of keys from  $Z$  contained in node  $j - 1$  and let  $P_{jmZ}^r = Pr.[jmZ = r]$ . Let  $PNCR_l = 1 - PCR_l$  denote the probability that  $j$  contains at least one key from keyset  $Z$ , for different values of  $\beta$ . Therefore, we have

$$PNCR_l = \sum_{\beta=0}^{k-1} P_{l,i,\beta}^{2P} \sum_{r=0}^{k-\beta} P_{jmZ}^r \left( \left\{ 1 - \frac{\binom{k-r}{fk}}{\binom{k}{fk}} \right\} + \frac{\binom{k-r}{fk}}{\binom{k}{fk}} \left\{ 1 - \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \right\} \right), \quad (3.6)$$

where the first term after the inner summation is the probability that at least one out of  $r$  keys is inherited by node  $j$  while the second term represents the complementary situation in which at least one key from keyset  $Z$  is obtained from the random keypool.

Next, using the fact that  $\frac{\binom{k-r}{fk}}{\binom{k}{fk}} \geq (1-f)^r$  and substituting in Equation 3.6, we have,

$$\begin{aligned}
PNCR_i^{2P} &\geq \sum_{\beta=0}^{k-1} P_{l,i,\beta}^{2P} \sum_{r=0}^{k-\beta} P_{jmZ}^r \left( 1 - (1-f)^r \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \right) \\
&= \sum_{\beta=0}^{k-1} P_{l,i,\beta}^{2P} \left( 1 - \left( \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \sum_{r=0}^{k-\beta} (1-f)^r P_{jmZ}^r \right) \right) \\
&= 1 - P_{l,i,k}^{2P} - \left( \sum_{\beta=0}^k P_{l,i,\beta}^{2P} \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \sum_{r=0}^{k-\beta} (1-f)^r P_{jmZ}^r \right). \tag{3.7}
\end{aligned}$$

Therefore we have

$$PCR_l^{2P} \leq P_{l,i,k}^{2P} + \sum_{\beta=0}^k P_{l,i,\beta}^{2P} \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \sum_{r=0}^{k-\beta} (1-f)^r P_{jmZ}^r, \quad l \leq \lceil \frac{j-i}{2} \rceil. \tag{3.8}$$

We obtain an efficient approximation for  $PCR_l^{2P}$  as follows: Let  $j-i=y$ ,  $j-l=m$ . Let  $p = k/L + B^y(1 - (k/L))$  if  $l < i$  and  $p = (1 - B^m)k/L$  if  $i < l \leq i + \lceil \frac{y}{2} \rceil$ , where  $B = (fL - k)/(L - k)$ . From lemma 2, the expected number of keys from  $Z$  present in node  $j-1$  is  $p(k-\beta)$ . For reasonably small values of  $k/L$ , we can therefore approximate the distribution of random variable  $jmZ$  by the standard Binomial distribution  $B(k-\beta, p)$  with the same mean  $p(k-\beta)$ . Therefore, we get

$$\begin{aligned}
PCR_l^{2P} &\leq P_{l,i,k}^{2P} + \sum_{\beta=0}^k P_{l,i,\beta}^{2P} \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \left( \sum_{r=0}^{k-\beta} (1-f)^r \binom{k-\beta}{r} p^r (1-p)^{k-\beta-r} \right) \\
&= P_{l,i,k}^{2P} + \sum_{\beta=0}^k P_{l,i,\beta}^{2P} \left( \frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \right) (1-fp)^{k-\beta} \quad l \leq \lceil \frac{j-i}{2} \rceil. \tag{3.9}
\end{aligned}$$

where  $p$  is defined as above.

By corollary 2,  $PCR_{i-t}^{2P} = PCR_{j+t}^{2P}$ ,  $t \geq 1$  and  $PCR_{i+t}^{2P} = PCR_{j-t}^{2P}$  for  $1 \leq t \leq \lceil (j-i)/2 \rceil$ . This defines  $PCR_l^{2P}$  for all values of  $l$  as specified in the statement of the proposition<sup>3</sup>.

Finally, using the fact that the probability of node  $j$  not containing any key from  $Z$  is  $\binom{L-k+\beta}{k} / \binom{L}{k}$  under random key predistribution, we derive  $PCR_l^{Rand}$  as:

---

<sup>3</sup>Henceforth, we will only use  $l \leq \lceil (j-i)/2 \rceil$  for the remaining propositions, using this symmetry.

$$PCR_i^{Rand} = P_{l,i,k}^{Rand} + \sum_{\beta=0}^k P_{l,i,\beta}^{Rand} \frac{\binom{L-k+\beta}{k}}{\binom{L}{k}} \quad (3.10)$$

We now consider two separate but related issues: First we determine values of  $f$  which minimize the probability of a given communication link  $(i, j)$  being compromised under the 2-Phase scheme. Given the security-performance tradeoffs, the user may desire a higher level of connectivity than provided by this optimal  $f$ . Therefore as an alternate performance metric, we determine values of  $f$  for which this probability is lower under 2-Phase key predistribution as opposed to Random key predistribution. ■

**Proposition 5**  $PCR_i$ , the probability of a given link  $(i, j)$  in the sensor net being compromised by capture of any node  $l$  is minimized by choosing an inheritance factor  $f$  that maximizes the expression

$$\begin{aligned} & x(1 - B^t)(1 - x(2f - f^2)) + f(1 - B^t)B^y(1 + fx - 2x)(1 - x) \\ & \quad \text{for } l = i - t \text{ and } l = j + t \quad t \geq 1 \\ & x(1 - B^t) [1 - x(2f - f^2)] - fB^{y-t}(1 + fx - 2x) \\ & \quad \text{for } l = i + t \text{ and } l = j - t \quad 1 \leq t \leq \lceil \frac{y}{2} \rceil \end{aligned}$$

where  $x = k/L$ ,  $j = i + y$  and  $B = \frac{L-k}{L-k}$ . If communicating nodes  $i$  and  $j$  are separated by a minimum distance (i.e.  $y > c$ , where  $c$  is a small constant), then  $f = \frac{1}{k}$  minimizes this link vulnerability.

**Proof:** We first obtain a simple approximation for Equation 3.9. From proposition 1, the expected number of keys in common between any two nodes  $i$  and  $l$  is given by  $p_{il} = k(x + B^{|i-l|}(1 - x))$ . Therefore for  $k \ll L$ , we can approximate the distribution of the number of common keys  $\beta$  between  $i$  and  $l$  by the Binomial  $B(k, p_{il}, \beta)$ . Now using  $\frac{\binom{L-2k+\beta}{k-fk}}{\binom{L-k}{k-fk}} \approx (\frac{1-2x+fx}{1-x})^{k-\beta}$ , we can rewrite Equation 3.9 for  $l \leq \lceil (j-i)/2 \rceil$  as

$$PCR_i^{2P} \leq P_{l,i,k}^{2P} + \begin{cases} \sum_{\beta=0}^k \binom{k}{\beta} (p_{il})^\beta \left( (1 - p_{il}) \left( \frac{1 - 2x + fx}{1 - x} \right) (1 - fx - f(1 - x)B^y) \right)^{k-\beta} & \text{if } l < i \\ \sum_{\beta=0}^k \binom{k}{\beta} (p_{il})^\beta \left( (1 - p_{il}) \left( \frac{1 - 2x + fx}{1 - x} \right) (1 - fx + fx B^{j-l}) \right)^{k-\beta} & \text{if } i < l < i + \lceil \frac{j-i}{2} \rceil \end{cases}$$

Substituting for  $p_{il}$  and further simplifying, we get

$$PCR_i^{2P} \leq P_{l,i,k}^{2P} + \begin{cases} \left( 1 - \left[ x(1 - B^t) (1 - x(2f - f^2)) + f(1 - B^t)B^y(1 + fx - 2x)(1 - x) \right] \right)^k \\ \text{for } l = i - t \text{ and } l = j + t \quad t \geq 1 \\ \left( 1 - x(1 - B^t) \left[ 1 - x(2f - f^2) - fB^{y-t}(1 + fx - 2x) \right] \right)^k \\ \text{for } l = i + t \text{ and } l = j - t \quad 1 \leq t \leq \lceil \frac{y}{2} \rceil \end{cases} \quad (3.11)$$

$PCR_i^{2P}$  is minimized by maximizing the inner term as stated in the proposition. When  $j - i \geq 4$  and  $k \ll L$ , this minimum value is obtained at  $f = 1/k$ . ■

**Proposition 6** *The probability of a given link being compromised by the capture of an arbitrary node  $l \neq i, j$  in the sensor net is lower under the 2-Phase scheme as compared to Random key predistribution for*

$$\frac{1}{k} \leq f \leq x \frac{2 - x}{1 + 2x}$$

where  $x = k/L$ .

**Proof:** We can express  $PCR_i^{Rand}$  in Equation 3.10 as

$$PCR_i^{Rand} = P_{l,i,k}^{Rand} + \left( 1 - \frac{k}{L} + \left( \frac{k}{L} \right)^2 \right)^k \quad (3.12)$$

by using the approximation  $\frac{\binom{L-k+\beta}{k}}{\binom{L}{k}} \approx (1 - \frac{k}{L})^{k-\beta}$  and also approximating the hypergeometric distribution of the number of common keys between nodes  $l$  and  $i$  by the Binomial  $B(k, \frac{k}{L})$  (assuming  $k \ll L$ ).

Now comparing Equations 3.12 and 3.11 using  $|l - i| = t$  and assuming  $j - i \geq 4$ , we have

$$\begin{aligned} PCR_i^{2P} \leq PCR_i^{Rand} \quad \text{iff} \quad & x(1 - B^t) \left[ 1 - x(2f - f^2) \right] \geq x(1 - x) \\ \implies & \left[ 1 - (1 - B^t)(2f - f^2) \right] x > B^t \end{aligned} \quad (3.13)$$

From the above expression, it can be seen that for reasonably large values of  $t$  (i.e. the captured node's LID is not too close to  $i$  or  $j$ ),  $PCR_i^{2P} < PCR_i^{Rand}$  for larger values of  $f < 1$ . However, the worst-case for link compromization  $PCR_i^{2P}$  occurs when  $t = 1$  i.e when nodes  $l = i + 1$  or  $l = j - 1$  are captured. Therefore substituting  $t = 1$  above and simplifying, we

get  $f < x(2 - x - 2f + 3f^2 - f^3)$  which implies  $\frac{1}{k} \leq f \leq x \frac{2-x}{1+2x}$ . Hence the 2-Phase scheme has lower vulnerability than the Random scheme for  $f$  upto  $2k/L$ . ■

From Equation 3.13, we can see that if the captured node is not too close to the communicating nodes (in terms of LID), then the 2-Phase scheme outperforms Random key predistribution for larger values of  $f$ , which in turn ensures higher connectivity. In particular, if the adversary is restricted to using knowledge of a captured node's keys within a small neighborhood such as its cluster, then we can further minimize link vulnerability to single-node capture by considering a modified 2-Phase scheme in which two neighboring nodes  $i$  and  $j$  with  $\geq q$  keys in common communicate only if there does not exist any other node  $l$  in the cluster such that  $i < l < j$ . Consider a sensor network with average node density  $M$ . Then the expected LID difference between node  $i$  and the nearest node (other than node  $j$ ) is  $N/M$ . We can therefore approximate link invulnerability to single node capture as follows:

**Proposition 7** *In a sensor network with average node density  $M$ , the probability that a given communication link  $(i, j)$  is invulnerable to single-node capture within its cluster is  $1 - (1 - PCR_{i-\frac{N}{M}}^{2P})^M$ .*

Simulation results in the next section illustrate link vulnerability within a cluster for different sensor network parameters.

Finally, for an average adversary with no specific knowledge about the network topology, the probability of capturing any node  $l \neq i, j$  is  $1/(N - 2)$ . The following proposition then directly follows proposition 6.

**Proposition 8** *The vulnerability metric  $VC$  of a given communication link in an  $N$ -node sensor network with parameters  $k$  and  $L$ , is lower if keys are predistributed using the 2-Phase scheme as compared to random key predistribution,  $\frac{1}{k} \leq f < x \frac{2-x}{1+2x}$ .*

### 3.9 Simulation Results

In this section, we describe some security and performance results based on simulations carried out on a 1000 node sensor network using a key pool  $L$  ranging from 8000-10000 keys. The per-node key space  $k$  varies from 40-150 keys. We have evaluated the 2-Phase key distribution scheme for  $f = 0.5$ . Nodes in the simulations are deployed in clusters as in LEACH [18] where the average node density (in a cluster) varies between 20 to 50 nodes.

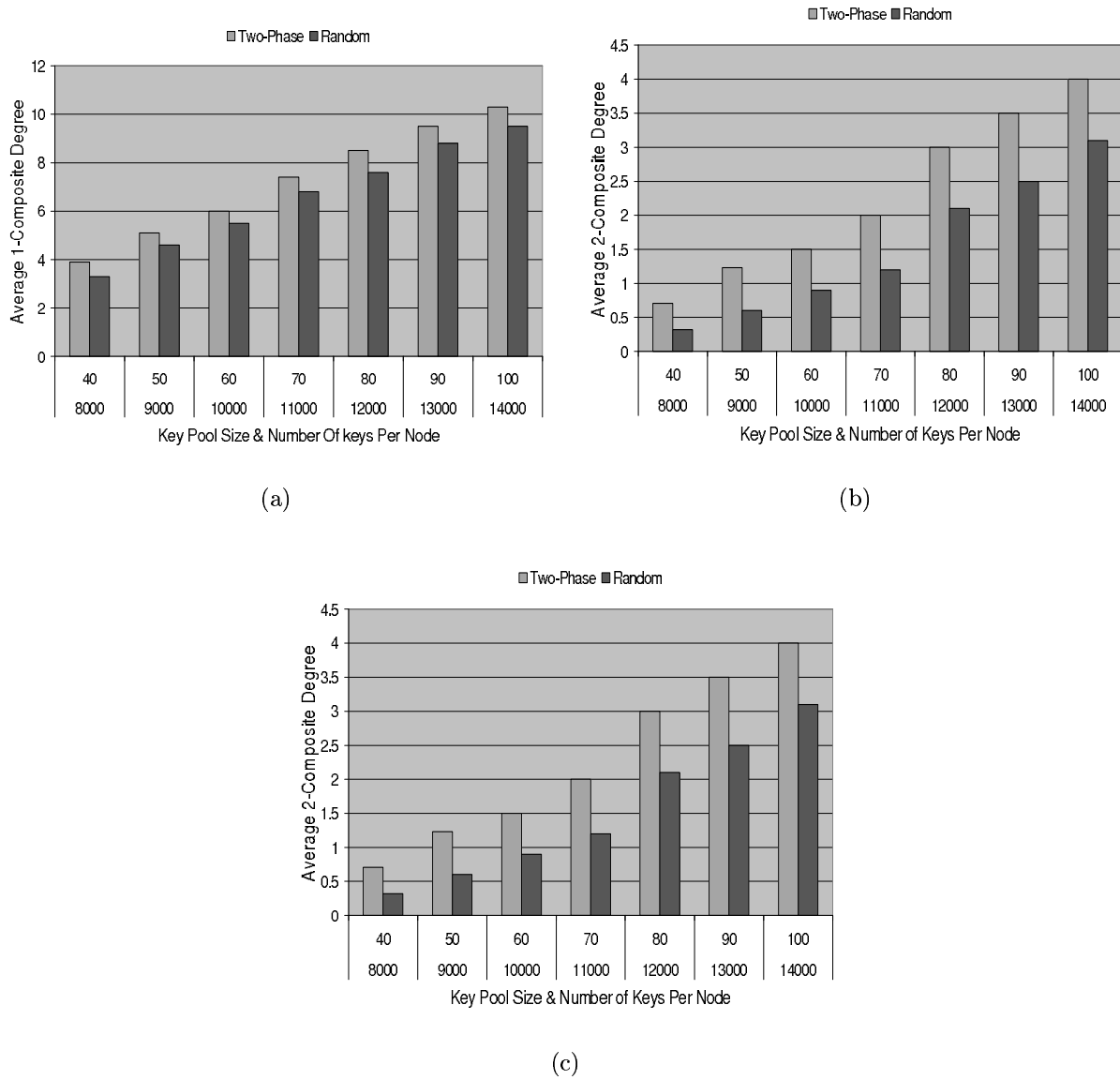


Figure 3.1: Average  $q$ -composite degree of a Node (a) 1-composite (b) 2-composite (c) 3-composite.

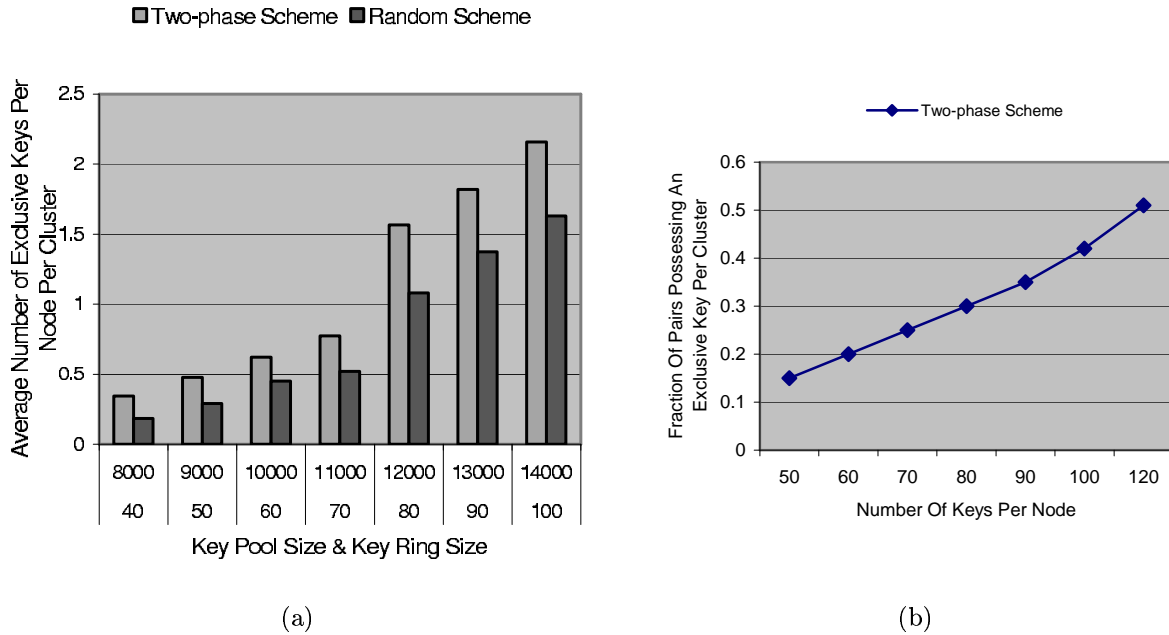


Figure 3.2: (a) Av. # exclusive keys per node pair. (b) Prob. a node pair has an exclusive key.

Figure 1 describe some  $q$ -composite network connectivity metrics while Figures 2–4 describe several sensornet security metrics.

Figure 3.1 describes the average  $q$ -composite degree of a node for different values of  $q$ . As can be seen clearly, the average degree is increasingly higher under 2-Phase and it outperforms the random key pre-distribution as  $q$  increases.

Figures 3.2–3.4 describe several sensornet security metrics. Figure 3.2(a) illustrates a measure of communication security (i.e invulnerability) by describing the average number of exclusive keys per pair of nodes in a cluster. This number is higher for nodes under 2-Phase than using the random scheme. Figure 3.2(b) measures the probability that a pair of nodes possesses at least one exclusive key under the 2-Phase key pre-distribution scheme. This probability rises sharply as  $k$ , the number of keys possessed by each node increases.

Figures 3.3 measure the vulnerability of communication links in a cluster under single as well as multiple node capture scenarios. As can be seen, the average number of links exposed to the adversary is lower under the 2-Phase scheme. The simulation results verify the analytical observations in Propositions 3.4 and 3.5 regarding link vulnerability. Lower link vulnerability for 2-Phase is explained by the fact that it is highly unlikely for captured

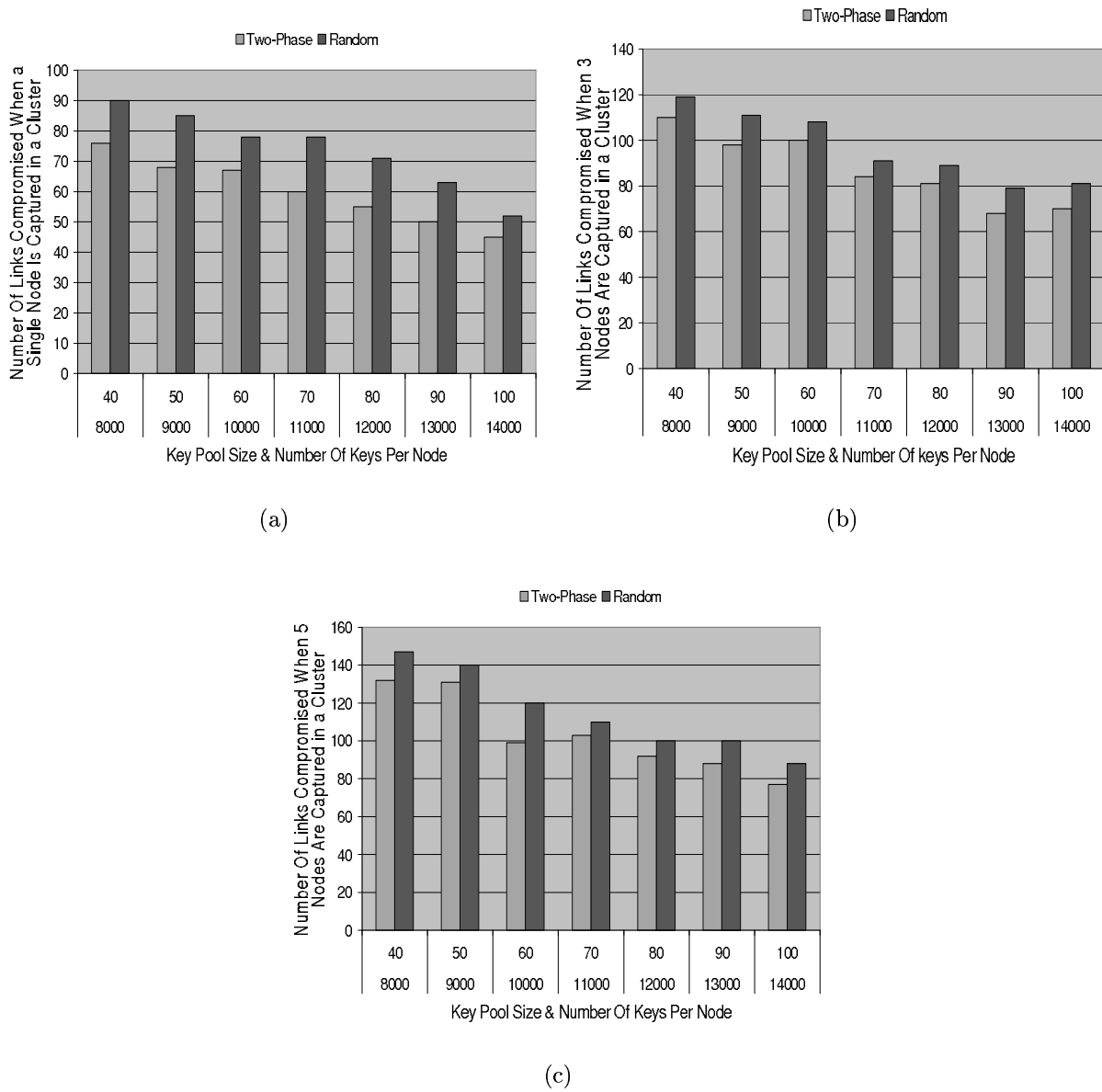
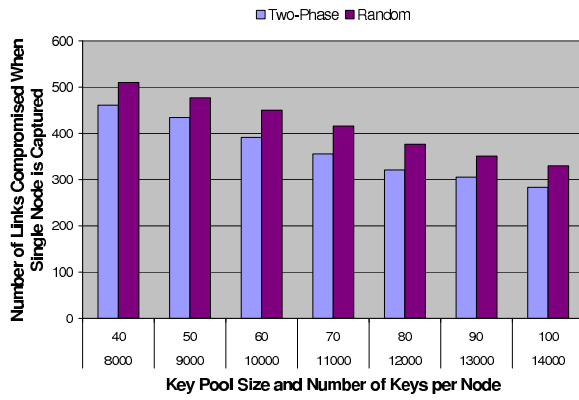
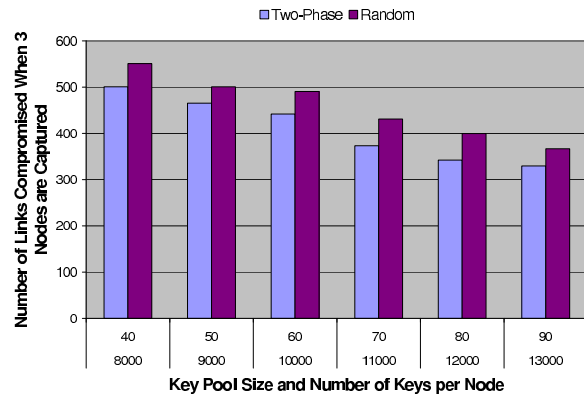


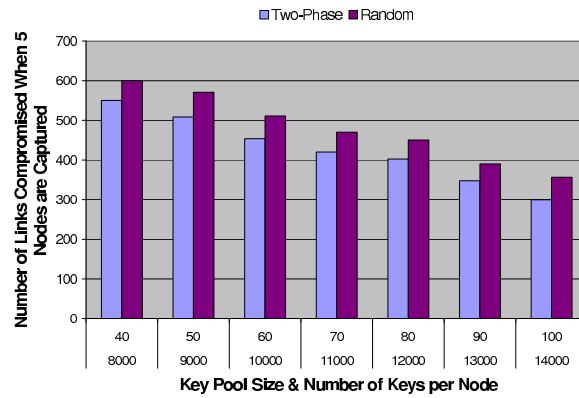
Figure 3.3: Average # links compromised in a cluster when (a) one (b) three (c) five nodes are captured.



(a)



(b)



(c)

Figure 3.4: # links compromised in the whole network when (a) one (b) three (c) five nodes are captured.

nodes to have an LID adjacent to the LIDs of the communicating nodes. We have also simulated network wide node capture scenario for various number of captured nodes. We have conducted the experiment on five networks of same size but with different configuration and different key pre-distributions. Figures 3.4 represent the average number of compromised links under different node capture scenarios, the average being taken over the five networks considered in the experiment. Clearly, the two-phase scheme outperforms the random scheme in this regard.

### 3.10 Implementation Issue: Creating Sorted Shared Key Lists

The security of a communication link strengthens with the exclusivity of the key(s) used for encryption on this link. For mutual communication each pair of nodes must therefore use keys shared among least number of nodes. During the shared key discovery phase, each node discovers its logical neighbors i.e., the neighbors with whom it shares at least one key. We propose the following metric to evaluate each shared key from this point of view.

Let  $k$  be a key shared between any two nodes  $i$  and  $j$  and let  $S_{ij}(k)$  denote the set of nodes in the neighborhood of  $i$  and  $j$  which share key  $k$ . Therefore, the eligibility of this key  $k$  with respect to the pair of nodes  $i$  and  $j$  is defined as:

$$E_{ij}(k) = \begin{cases} 1 & \text{if } S_{ij}(k) = \phi \\ \frac{1}{|S_{ij}(k)|} & \text{otherwise} \end{cases} \quad (3.14)$$

The higher the value of  $E_{ij}(k)$ , the better is the key  $k$  for communication between  $i$  and  $j$ . During the shared key discovery phase, each node broadcasts the list of identifiers of the keys it possesses. Each node then create a separate list of shared keys for each of its neighbors sorted according to their eligibility values. The most eligible key should be used for communication until it is revoked.

### 3.11 Conclusion

Efficient pre-distribution of keys to sensor nodes is a very important issue for secure communication in sensor networks. Connectivity and resiliency to enemy attacks must be traded off very carefully. In this paper, we present an analytical framework with several quantitative metrics for evaluating key predistribution schemes and determining their security-performance tradeoff. We also present a 2-Phased key predistribution scheme based on a combination of inheritance and randomness which is proved to have better tradeoffs.

# Chapter 4

## Secure Data Aggregation

### 4.1 Introduction

Data aggregation is one of the most important data-centric mechanisms/operations of a sensor network. A set of special nodes called *aggregators*, desirably less energy-constrained, perform in-network data aggregation and get rid of redundant data, thereby enabling energy-efficient information flow to a querying remote user [67], [68]. Data aggregation has enormous impact on energy savings of a severely energy-constrained wireless sensor network with many-to-one redundant low-rate data flow. The pioneering work on impact of data aggregation [67] shows that data-centric routing offers significant performance gains across a wide range of operational scenarios. However, the task of data aggregation becomes challenging when sensors and aggregators are deployed in a harsh or hostile environment and become subject to mechanical fault or physical tampering. A single compromised sensor device can render the whole network useless, or worse, mislead the user into trusting a false reading causing catastrophic consequences in many important applications. For example, an adversary could make a battlefield surveillance network fail to report an intrusion without creating any suspicion. The foremost requirement of a data processing mechanism thus is that it must be fault tolerant and resilient to enemy attacks.

In this chapter, we address the problem of enabling correct information aggregation, given that a fraction of the aggregators and ordinary sensors might be faulty or compromised. Our contributions regarding this problem are as follows:

- Previous works on this problem provides a solution by unrealistically limiting the number nodes that an adversary can compromise. We provide a more general solution which does not require this assumption.

- We provide an analytical model of computing data accuracy based on *spatial correlation* of data values.
- We provide a novel algorithm which computes a *weighted aggregate* of data by attaching less weights to the sensor readings which are more likely to be wrong according to the proposed data accuracy model.
- We show by simulation that the weighted aggregate is more accurate than a simple aggregate.
- We consider the problem of data aggregation when an aggregator is corrupted. We propose several solutions based on statistical estimation of parameters in a multi-aggregator set-up.

## 4.2 Literature Review on Data Aggregation and Security

Previous research work on data aggregation mainly focused on its impact on the performance and energy savings of wireless sensor networks [67, 68, 75, 76] assuming that there is no security threats. A few work have been done on security of data aggregation [77, 69]. In this section, we review the works specifically done for secure data aggregation and the fault tolerant aggregation protocols.

### 4.2.1 Secure Aggregation for Wireless Network

The first paper [77] written by Lingxuan Hu and David Evanson on secure aggregation was published in 2003. In this paper, the authors focus on an adversary whose objective is to provide false information to the user. They propose a secure aggregation protocol based on several basic assumptions as stated below:

- The base station trusts results from a sensor network.
- The base station has enough capacity to broadcast messages to all nodes directly. Sensor devices having limited energy and transmission range, they can only communicate with nearby nodes .
- Low-level network mechanisms ensure reliable message delivery.

- The distance between a typical node and a base station is likely to be very large in terms of hops.
- The network is so dense that each node has several other nodes in one-hop distance.
- Before deployment, each node can establish secret key with the base station.
- Each node has a unique node ID.
- A secure self-organizing protocol is used to form a tree-based routing hierarchy where each node has an immediate parent.

Based on the above assumptions, the proposed protocol exploits two key ideas: *delayed aggregation* and *delayed authentication*. Instead of aggregating messages at the immediate next hop, messages are forwarded unchanged over the first hop and then aggregated at the second hop. While this increases the transmission costs, it guarantees integrity for networks where two consecutive nodes are not compromised. A delayed authentication not only saves energy resources enables but also enables usage of symmetric keys which are revealed to the authenticator after the time delay has expired.

The key steps of the protocol are as follows:

- Each leaf node transmits its reading to its parent. A raw message includes the node data reading and the node ID. A message authentication code (MAC) is also included with the message. The code is encrypted using the secret key shared between the sender node and the base station.
- The parent node saves the message until it receives the secret key from the base station and verifies the MAC. It raises an alarm if the MAC does not match.
- Message aggregation is performed in each intermediate step. Nodes wait for a specified time to receive messages from their children, and then retransmit the messages and MACs they receive directly from immediate children. Each child can contribute to at most one reading in each time.
- Nodes aggregate the data they receive from their grandchildren (via their children) and transmit the MAC of the aggregation value.
- Delayed aggregation ensures that an adversary who obtains a key from a compromised node cannot tamper with many sensor readings.

- After a stage of messages arrives at the base station, the base station reveals the temporary node keys along with a MAC generated using base stations current  $\mu$ TESLA key.
- Once the key is revealed, nodes advance to the next temporary node key. After this, the  $\mu$ TESLA key is revealed to enable authentication, and the base station advances to the next key in the chain.

Thus, the proposed secure aggregation protocol provides security at the expense of little delay and routing cost. However, the main drawback of this protocol is that it provides security against one node compromise. The protocol becomes vulnerable if a parent and a child node in the hierarchy are compromised at the same time.

#### 4.2.2 SIA: Secure Information Aggregation in Sensor Networks

This paper [69] presents a framework for secure data aggregation under a *polynomially bounded stealthy attacker* who can arbitrarily change the data value sent by a compromised sensor/aggregator. The authors study that there is a tradeoff between the communication expense and 100 percent accurate information sent to the base station. Assuming the the communication between an aggregator and the base station, the authors propose a protocol with sublinear communication complexity compromising some degree of accuracy. In short, the proposed protocol makes a base station accept an approximately correct aggregated data value.

The authors propose *aggregate-commit-prove* approach: an aggregator replies to the user with the aggregation result together with a commitment to the collection of data. The user and the aggregator then perform efficient interactive proofs such that the user will be able to verify the correctness of the results (or detect cheating) with high probability. A brief overview of the steps of the solution is given below:

- The aggregator aggregates data received from sensors after verifying the authentication of the message using a key shared between the aggregator and that sensor.
- The aggregator commits to the aggregated data constructing a Merkle hash tree. The tree starts with the original data values as leaf nodes. Each internal node in the binary hash tree is a hash of the concatenation its two children values. The root of this tree, called the *commitment* is sent to the user.
- As the user receives the commitment, it starts an interactive proof protocol to check whether the commitment is a good representation of the data or the aggregator is cheating. The aggregator needs to draw some random samples again from the network.

This protocol has the following drawbacks:

- Some compromised sensors can report wrong values that will affect the aggregation result. The authors calculate the bound on the deviation these corrupted sensors can cause assuming that at most a certain number of sensors can be compromised. However, this assumption is not realistic.
- The proposed protocol can detect an error with high probability if the error is within a small bound, based on the assumption that a corrupted aggregator will inject only a slight error in the original aggregated result. This assumption, too, may be far from reality in case of some applications.
- Although an aggregator can detect sensor impersonation using message authentication key, it cannot detect a flawed data from a corrupt sensor.

### 4.3 Attack Model and Security Goals

An attacker can perform a wide variety of attacks which are jeopardizes the communication in any part of a sensor network. For example, an adversary can compromise any node, be it an ordinary sensor or a base station, and perform denial of service (DoS) attack. The only remedy for a DoS attack is to detect a compromised node and stop communicating with it. A different kind of attack is a *stealthy attack* in which an adversary makes a compromised node significantly alter the data so that a user accepts a false result. For applications in battlefields or in disaster relief operations, a false data can create disastrous consequences. What makes a stealthy attack worse is that it is impossible to detect a stealthy attack with 100 percent accuracy. However, very little work have been done on this problem [77, 69]. While [77] provides very little security against this attack at the expense of some communication and delay overhead, [69] provides solutions based on unrealistic assumptions also with some communication overhead.

We focus on the same problem setting considered in [69]. We assume the Byzantine fault model where a compromised node is under the full control of the attacker and the adversary can misbehave in any arbitrary way. However, unlike [69], we do not pose any limitation on the adversary's computational resources and the fraction of nodes that it can corrupt. In this setup, we consider stealthy attacks where the attackers objective is to make a user accept significantly incorrect aggregation results without being able to detect it. Note

that in a severely resource-constrained wireless sensor network, it is practically impossible to detect an incorrect result with perfect accuracy since it involves huge computational and communication cost. Therefore, the goal is to provide some security mechanism which enables a user to detect any misbehavior with a considerably high degree of accuracy at a very little expense.

## 4.4 Problem Setup and Statement

We consider the following problem setting. A large number of sensors are deployed in an area distant from a *user/server* which would like to get information about that area covered by the sensors. However, sensors are typically simple, low-powered devices which can communicate within a small range of their location and are thus unable to send information directly to the home server. In such a situation, one or more resource enhanced base stations are used as an intermediary between the home server and the ordinary sensors. We assume a cluster based network architecture as described in [18]. Each cluster has at least one clusterhead/gateway node which is less energy-constrained, computationally more powerful and directly communicate with the base station. These clusterheads collect and aggregate data from the corresponding clusters and send to the base station which in turn sends it to the user through internet. In this paper, we consider multi-aggregator setup i.e., there are more than one aggregators for a particular cluster. They all collect the same information from all the sensors, process data and send to the base station. This enables the user to compare the results and select the correct result with high probability. Ideally, all aggregators from the same cluster should send same data value. When the base station receives different data values, it suspects that a corruption has taken place. It is highly unlikely that all the aggregators will be compromised/faulty at the same time. Hence, a user is guaranteed with high probability to receive at least one correct data value.

In order to enable secure and authentic communication, all data must be encrypted. However, sensor networks being stringently resource constrained, public key cryptography cannot be implemented. In this paper, we assume that each sensor has a unique identifier and shares a separate private/secret key with the base station and all the aggregators.

We divide the problem into two parts: a) a fraction of sensors are faulty/corrupt and an aggregator needs to detect inaccurate information; b) a fraction of aggregators are faulty/corrupt and the user needs to detect the corruption and accept only the true result with high probability. For the first part of the problem, an aggregator estimates accuracy of each sensor reading by exploiting the fact that sensor readings are spatially correlated and computes weighted data aggregation. In the second part of the problem, if the user receives different data values from different aggregators, it detects a corruption and collects a random

sample directly from the sensors. Based on that sample, it performs a statistical testing of hypothesis [73] and accepts the correct result with high probability. In this problem setup, our contributions are the following:

- We propose a probabilistic data accuracy model which computes accuracy of each sensor reading exploiting the fact that sensor readings are spatially correlated. This generalized model is appropriate for any real number data.
- We propose several statistical approaches which enable a user to accept the true aggregated value from an aggregator with correct readings with high probability.

## 4.5 Weighted Data Aggregation

An aggregator aggregates raw data collected from every sensor node of a region and sends the data directly to a home server/user. A fraction of these sensors can be corrupt or faulty and thus can produce wrong information. The objective of the aggregator is to identify and discard these wrong information values as much as possible so that the aggregated data can be a good/nearly correct representative of the data collected from the whole region. It is impossible for the aggregator to identify corrupt/faulty sensors or incorrect data values every time it receives a set of data due to the following reasons:

- The set of faulty/corrupt sensors may change over time.
- In order to verify whether a data from a particular sensor is correct or not, numerous messages should be exchanged between the aggregator and that particular sensor. A sensor network being stringently energy constrained, verification of every single information is impossible.

While it is impossible to differentiate between accurate and inaccurate data values, an aggregator can estimate degree of accuracy of an information. However, methods of such estimation are solely application-specific.

In this paper, we address this problem for a set of specific applications with the following characteristics:

- Data values are real numbers.
- Data values are spatially correlated.

Example of such applications are numerous - weather related applications in which temperature, humidity, air pressure etc are measured in regions where human intervention is

not possible, health related applications e.g., glucose monitor, organ monitor, general health monitor etc. We propose a solution in which the aggregator exploits the spatial correlation among data values to estimate accuracy of each data value. According to estimated accuracy of each data value, the aggregator attaches weight to them and finally computes weighted data aggregation. We now finally describe our model below.

We propose a probabilistic data accuracy model based on the following assumptions:

- Let  $j$  be the point which is covered by a sensor node  $s_j$ . The true/correct reading of that point is denoted by  $t_j$ . Then the data value sent by node  $s_j$  can be expressed as:  $r_j = t_j + e_j$ , where  $e_j$  is the amount of error in the reading. For all sensors,  $e_j$ s are independently distributed variables. Distribution of any  $e_j$  depends on whether the node  $s_j$  is faulty or not.
- A node  $s_j$  is faulty/corrupt with probability  $q$ . The value of  $q$  is discussed later.
- If a node is not faulty/corrupt,  $e_j = 0$  with probability 1.
- We consider two different models - fault-tolerance model and adversary attack model. Our proposed model and data aggregation algorithm works assuming only one of these two models.
- In a fault-tolerance model, if a sensor  $s_j$  is faulty,  $e_j$  is distributed normally with mean 0 and variance 1.
- If we consider an attack model where an adversary captures a node and sends false reading to the aggregator, then  $e_j$  is distributed uniformly within  $a$  and  $b$ , where  $a$  and  $b$  are parameters specified by the adversary.
- A reading  $r_j$  is considered accurate if  $t_j - \eta \leq r_j \leq t_j + \eta$ .  $\eta$  is an application specific parameter and is determined a priori. Clearly, the probability that a sensor reading is accurate is given by:  $P[t_j - \eta \leq r_j \leq t_j + \eta]$  i.e.,  $P[|e_j| \leq \eta]$ . We denote this probability as  $p$ .  $p$  can be expressed as follows:

$$\begin{aligned}
 p &= P[|e_j| \leq \eta] \\
 &= P[|e_j| \leq \eta \mid s_j \text{ is not faulty/corrupt}](1 - q) + P[|e_j| \leq \eta \mid s_j \text{ is faulty/corrupt}]q \\
 &= (1 - q) + qP[|e_j| \leq \eta]
 \end{aligned} \tag{4.1}$$

$P[|e_j| \leq \eta]$  is evaluated according to either the fault-tolerance model or the adversary attack model.

- We define *spatial neighborhood* of a sensor node  $s_i$  as the set of nodes who are within a range of predetermined radius  $R$  so that the true readings should not vary from the reading of  $s_i$  by  $\varepsilon$ .  $R$  should be so chosen that the readings of two sensors within this neighborhood are highly correlated. Clearly,  $\varepsilon$  is a function of  $R$ .

Now, we define accuracy model of a sensor reading which exploits the spatial correlation defined above.

Let  $s_i$  and  $s_j$  be two spatial neighbors. The absolute difference between the value of information sent by these two nodes is denoted by  $D(s_i, s_j)$ . We define *conditional accuracy* of  $s_i$  with respect to  $s_j$  as a fraction between 0 and 1 which expresses the relative accuracy of  $s_i$  given accuracy of node  $s_j$ . The conditional accuracy is defined as follows:

$$Acc(s_i|s_j \text{ is accurate}) = \begin{cases} 1 & \text{if } D(s_i, s_j) < \varepsilon \\ e^{-\alpha D(s_i, s_j)} & \text{if } \varepsilon \leq D(s_i, s_j) \end{cases} \quad (4.2)$$

Clearly, the node  $s_j$  being accurate, accuracy of  $s_i$  reduces exponentially as the difference between the readings of  $s_i$  and  $s_j$  increases. Accuracy of  $s_i$  becomes 0 if  $D(s_i, s_j)$  is considerably large.  $\alpha$  is a positive constant that indicates how much the accuracy of  $s_i$  is affected by  $D(s_i, s_j)$ .

In a similar manner,

$$Acc(s_i|s_j \text{ is inaccurate}) = \begin{cases} \varepsilon & \text{if } D(s_i, s_j) < \varepsilon \\ \gamma & \text{if } \varepsilon \leq D(s_i, s_j) \end{cases} \quad (4.3)$$

In this case, the reading of  $s_j$  being inaccurate,  $s_i$  is also inaccurate with a high probability if  $D(s_i, s_j)$  is very small. On the other hand, if  $D(s_i, s_j)$  is not so small, it is difficult to predict accuracy of  $s_i$ . Therefore, we attach a constant quantity  $\gamma$  as  $s_i$ 's accuracy in such a way that  $0 < \varepsilon < \gamma < 1$ .

The expected accuracy of a node  $s_i$  with respect to node  $s_j$  is:

$$A_j(s_i) = \begin{cases} (1 - \varepsilon)p + \varepsilon & \text{if } D(s_i, s_j) < \varepsilon \\ pe^{-\alpha D(s_i, s_j)} + (1 - p)\gamma & \text{if } D(s_i, s_j) \geq \varepsilon \end{cases} \quad (4.4)$$

The neighborhood  $N(i)$  of a sensor  $i$  is divided into two exclusive and exhaustive groups as:

$$\begin{aligned} N_1(i) &= \{s_j : D(s_i, s_j) < \varepsilon\} \\ N_2(i) &= \{s_j : D(s_i, s_j) \geq \varepsilon\} \end{aligned}$$

The *accuracy index*  $AI(i)$  of sensor  $i$  is defined as:

$$AI(i) = \sum_{j \in N_1(i)} A_j(s_i) \quad (4.5)$$

Note that if the set  $N_1(i)$  is empty, then  $AI(i) = 0$ .

The *inaccuracy index*  $II(i)$  of sensor  $i$  is defined as:

$$II(i) = \sum_{j \in N_2(i)} A_j(s_i) \quad (4.6)$$

Note that if the set  $N_2(i)$  is empty, then  $II(i) = 0$ .

An aggregator assigns weight to each sensor reading according to the following rule:

- A sensor reading will be given weight  $W(i) = \frac{AI(i)}{AI(i)+II(i)}$ .

The aggregator thus computes weighted average of the received data. It can also compute weighted median, weighted maximum or minimum.

We observe the following facts:

- The weight of a sensor reading lies between 0 and 1 and is proportional to its accuracy index. Therefore, a sensor reading is likely to have higher weight whenever its reading will be close to most of its spatial neighborhood readings.
- In an extreme case when no sensor is faulty/corrupt in a spatial neighborhood, all sensor readings will be the same, thereby making the inaccuracy index of a sensor 0 and its weight 1. However, in the other extreme case where all the sensors in a spatial neighborhood is faulty/corrupt, sensor readings are highly likely to vary in a wide range. In this case the set  $N_2(i)$  of a sensor  $s_i$  is highly likely to be larger than  $N_1(i)$ . Thus, the inaccuracy index is likely to be higher, thereby making the weight of  $s_i$  less than 0.5.
- In case of an enemy attack, the weighted data aggregation method will work perfectly well if an adversary chooses a sensor node at random to compromise i.e., if the corrupt sensors are dispersed uniformly over the sensor network.

## 4.6 Error Analysis

We consider the spatial neighborhood  $N(i)$  of a sensor  $i$  with size  $n$ . Let the size of  $N_1(i)$  be denoted as  $x$ . Then the size of  $N_2(i)$  is  $n - x$ . Each sensor is accurate with a probability  $p$ .

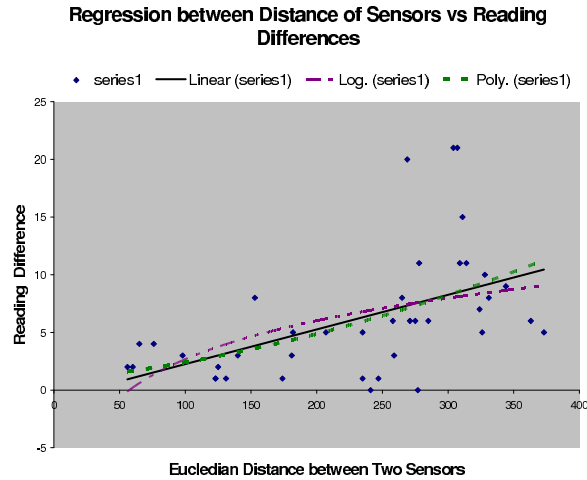


Figure 4.1: Different Regression Lines Fitted Through Reading Difference Data Plotted Against Sensor Distance

For simplicity of computation, we assume that accuracy of a sensor reading is independent of others. In other words, a sensor is faulty/corrupt independent of others. Therefore, the number of accurate sensors in  $N(i)$  i.e.,  $x$  follows Binomial distribution with parameters  $n$  and  $p$ . In this paper, we consider two extreme accuracy errors i.e., a sensor reading is accurate but  $x = 0$  and a sensor reading is inaccurate but  $x = n$ . Let the errors called as Type 1 error and Type 2 error respectively. Let the probabilities of these two errors be  $P_1$  and  $P_2$  respectively. Then,

$$P_1 = p(1-p)^n \quad (4.7)$$

Similarly,

$$P_2 = (1-p)p^n \quad (4.8)$$

Ideally, probabilities of these two errors should be as small as possible. Since the probabilities depend on the tolerance limit  $\eta$ ,  $\eta$  should be chosen in such a way that both  $P_1$  and  $P_2$  are minimum. Unfortunately, the value of  $\eta$  that minimizes one probability maximizes the other. Therefore, we need to select  $\eta$  in such a way that both  $P_1$  and  $P_2$  are reasonably small.

## 4.7 Selection of Parameters

### 4.7.1 Determination of $q$

In the model presented above, the most important parameter is  $q$ , the probability of a node being faulty/corrupted. This probability can be estimated in many ways. In case of a fault-tolerance model, we can assume that a sensor can be either faulty or faultless and therefore  $q = 0.5$ . Similarly, in the adversary attack model, a sensor node can be assumed to be equally likely to be captured or not captured by an adversary. Hence,  $q = 0.5$ . We propose another iterative mechanism for determining  $q$  which can be applied in both the fault-tolerance model and the adversary-attack model.

- Let  $q_t$  be the probability of a node being faulty/compromised at time  $t$ . Let  $N_t$  denote total number of nodes sending information to an aggregator at time  $t$ . Let  $M_t$  be the number of nodes at time  $t$  for which accuracy index is higher than inaccuracy index. The accuracy and inaccuracy indices at time  $t$  are calculated based on  $q_{t-1}$ . We assume that  $q_0 = 0.5$ . Then,  $q_t = \frac{M_t}{N_t}$ .

### 4.7.2 Determination of $\epsilon$

For positively spatially correlated data, reading difference of two neighboring nodes cannot exceed a certain limit. The parameter  $\epsilon$  is the index of that limit. The value of  $\epsilon$  has to be set a priori. However, in order to set a value for  $\epsilon$ , we must first quantify the correlation between reading difference of two sensors and their physical distance. One simple way to determine this is to find a regression curve of the two variables reading difference and physical distance. Here, the question arises about the nature of the regression. Note that there are two variables, the distance between two sensors, say  $Ds$ , and the difference of reading between these two sensors, say  $Df$ .  $Df$  is a function of  $Ds$ . This function may not be linear. But in practical problems, often the true regression curve is approximated by a straight line using *least-square principle*. This means that we choose among all straight lines in the  $(x, y)$  plane the one for which the sum of square deviation is minimum. Let the regression line be :  $Df = \omega + \phi Ds$ . There are two ways to determine the constants  $\omega$  and  $\phi$ . One way is to collect sample data beforehand (not through sensors) and fit a line through the sample data. Another way is to determine the bivariate probability distribution of the two variables i.e., reading difference and physical distance, and calculate the regression coefficients.

Regression Line Using Sample Data: In order to determine the constants  $\omega$  and  $\phi$ , we collect a sample of bivariate data  $Df_i, Ds_i, i = 1, 2, \dots, M$ . Then we minimize the quantity

$\sum_i (Df_i - \omega + \phi Ds)^2$  by equating this to zero. We get the following solution:

$$\omega = \bar{Df} - r \cdot \frac{s_{Df}}{s_{Ds}} \bar{Ds} \quad (4.9)$$

$$\phi = r \cdot \frac{s_{Df}}{s_{Ds}} \quad (4.10)$$

$\bar{Df}$  and  $\bar{Ds}$  are sample averages of  $Df$  and  $Ds$ .  $r$  is the correlation coefficient between  $Df$  and  $Ds$  and  $s_{Df}$  and  $s_{Ds}$  are the standard deviations of  $Ds$  and  $Df$  respectively. Figure 4.1 shows a regression line for temperature data.

Regression Line Using Bivariate Probability Distribution: In this case, the mean square deviation is defined as  $E[Df - \omega + \phi Ds]^2$ . The solution for  $\omega$  and  $\phi$  are given as:

$$\omega = \mu_{Df} - \rho \cdot \frac{\sigma_{Df}}{\sigma_{Ds}} \mu_{Ds} \quad (4.11)$$

$$\phi = \rho \cdot \frac{\sigma_{Df}}{\sigma_{Ds}} \quad (4.12)$$

$\mu_{Df}$  and  $\mu_{Ds}$  are expectations of the variables  $Df$  and  $Ds$  respectively. Similarly,  $\sigma_{Df}$  and  $\sigma_{Ds}$  are the population standard deviations.  $\rho$  is the correlation coefficient of the joint probability distribution of  $Df$  and  $Ds$ .

Suppose, the radius of the spatial neighborhood of a sensor is  $R$ . Then, the maximum reading difference between two sensors in a neighborhood can be set as:  $\epsilon = \omega + \phi R$ .

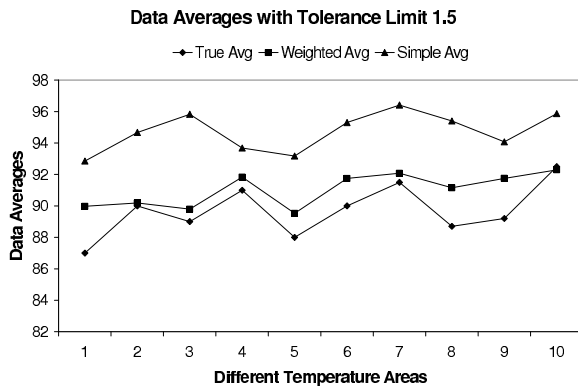
The other two parameters  $\alpha$  and  $\gamma$  are empirical and can be determined based on specific applications and through experimentation.

## 4.8 Performance Evaluation

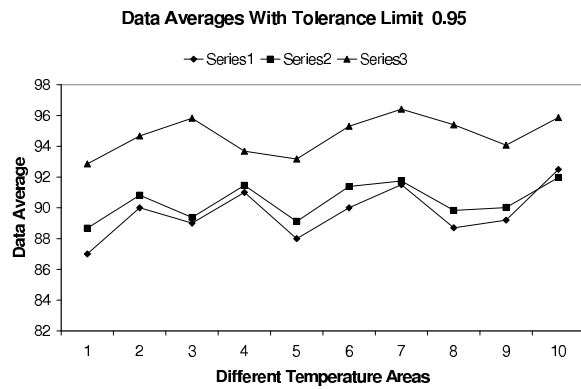
The objective of the weighted data aggregation algorithm is to remove errors as much as possible so that the aggregated data is nearly accurate i.e., the aggregated data does not vary too much from the true data values. In order to testify effectiveness of our proposed algorithm, we *data average* as the test metric. We compare the weighted average of the reported data with the simple (without weight) average of the reported data and the true average (i.e., the average data value of the correct data). In the next subsection, we describe our experimental setup.

### 4.8.1 Experimental Setup

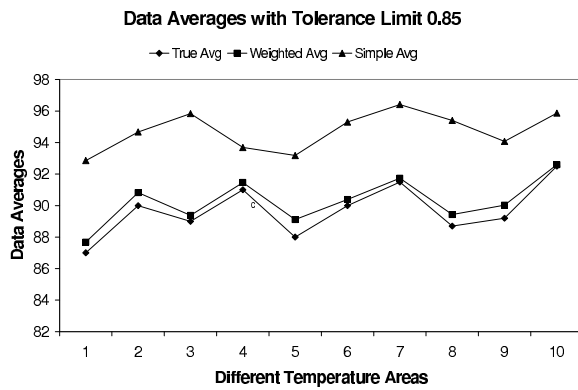
For our experiment, we consider ten adjacent cities of Louisiana. We collect temperature data for a particular day and time for all those cities. In the simulation setup, we deploy



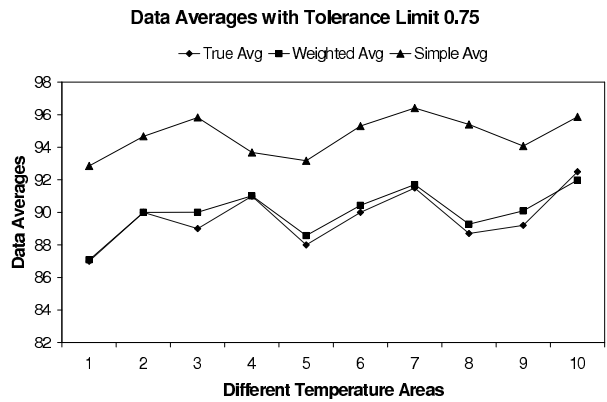
(a) Tolerance Limit=1.5



(b) Tolerance Limit=0.95

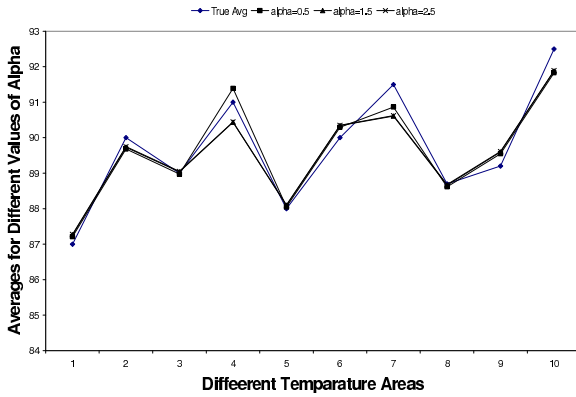


(c) Tolerance Limit=0.85

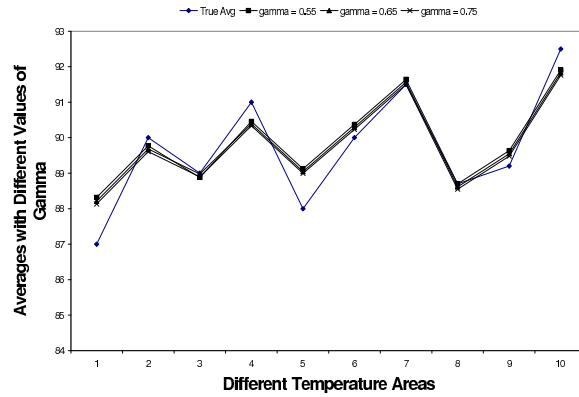


(d) Tolerance Limit=0.75

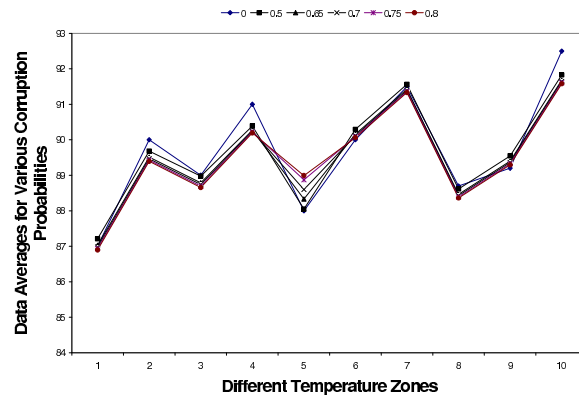
Figure 4.2: Simulation results



(a)



(b)



(c)

Figure 4.3: Trend of weighted data aggregation for different values of (a)  $\alpha$ , (b)  $\gamma$  and (c)  $p$

hundred sensor nodes in each of the cities. If a sensor is uncorrupt, it sends the correct temperature value to the aggregator. If a sensor is corrupt, it sends a faulty reading to the aggregator. If the aggregator finds a reading unusually high or low, it immediately suspects corruption and discards the reading. Therefore, the error injected in a sensor reading must have some upper and lower limits. In this particular setup, the range of temperature data in the ten cities is  $87^{\circ}$  to  $93^{\circ}$ . We conduct our experiment with error limit  $95 - 110$ . Varying the tolerance limit, we show that for a fixed set of values of  $\alpha$ ,  $\beta$  and  $\gamma$ , the weighted average converges to the true average as the tolerance limit reduces. Varying the tolerance limit, we show in figures 2 that for a fixed set of values of  $\alpha$ ,  $\beta$  and  $\gamma$ , the weighted average converges to the true average as the tolerance limit reduces. In this simulation setup, each sensor is corrupt with probability 0.5. We also study effects of different parameter values on the weighted average. Figures 3(a) and (b) show that the weighted average does not vary significantly with different values of  $\alpha$  and  $\gamma$ . Figure 3(c) shows that although different values of probability of corruption do not produce significant changes in the weighted average, the weighted average shifts slightly from the original average as the probability of corruption increases.

## 4.9 Compromised Aggregators: Problem and Solutions

An aggregator node sends aggregated data to a user/home server. Ideally, the data sent by all the aggregators should match since each of them collected information on the same set of sensors. However, due to fault or compromisation, aggregators can send wrong data. Therefore, the home server may receive different data from different aggregators. In this paper, we do not impose any assumption on the maximum number of aggregators that can be corrupted or faulty. Thus, it becomes impossible for the server to find out the correct data deterministically. We consider two cases: - collaborative model and non-collaborative model.

1. Collaborative model is appropriate only in case of an adversary attack where the compromised aggregators collaborate and send similar results to the server. In this case, the server will receive two different data - one from the group of good/uncompromised aggregators and the other from the group of compromised aggregators. Since the majority group can be corrupted, the server needs to test which of the two data is correct with higher probability.
2. A non-collaborative model can be applied if the aggregators are faulty or if the adversary is not so powerful as to make the compromised aggregators collaborate. Thus, in this model, faulty/corrupt sensors send different results to the server. In this case,

the server receives a set of different results and it needs to test which one of them is probabilistically correct. This case is more general.

In order to solve these problems, we make following assumptions:

- Sensor readings are real numbers.
- We consider applications where some natural phenomenon is detected the sensors. Therefore, it is realistic to assume that a sensor reading is a Gaussian variable with parameters  $\mu$  and  $\sigma$ ,  $\mu$  and  $\sigma$  being unknown.
- Each aggregator sends *weighted average* of data values to the user.

## 4.10 Solutions: Our Approach

A user receives a set of aggregated data from a set of aggregators. Ideally, all data should be the same. However, when a fraction of aggregators are faulty/corrupted, the user receives a set of different data values and detects corruption. It then requests the aggregators to collect a random sample of size  $n$  from the ordinary sensors and send  $n$  raw data to the user. Note that this time sensors are required to send data encrypted with the keys they share with the user. Therefore, in case of an adversary attack model, corrupt aggregators cannot alter this data even if the data is sent to the user via them.

Let  $x_1, x_2, \dots, x_n$  be the random sample of size  $n$  collected by the user. Let  $\mu_1, \mu_2, \dots, \mu_k$  be the distinct aggregated data sent by  $m$  aggregators,  $1 < k < m$ .

### 4.10.1 Bayes Estimate

This approach assumes a fixed a priori probability  $p_f$  of an aggregator being faulty. Using this probability and the received data from all the aggregators, the user takes a decision as to which data is to accept.

The user/home server receives several weighted averages of data values of a particular region from several aggregators. Let  $\mu_1, \mu_2, \dots, \mu_k$  be the distinct aggregated data sent by  $m$  aggregators,  $1 < k < m$ , with frequencies  $f_1, f_2, \dots, f_k$  respectively.

Since faults in aggregators occur independently, we estimate probability distribution of the parameter  $\mu$  using the observed frequencies as follows:

$$P[\mu = \mu_j] = (1 - p_f)^{f_j} p_f^{k - f_j} \quad (4.13)$$

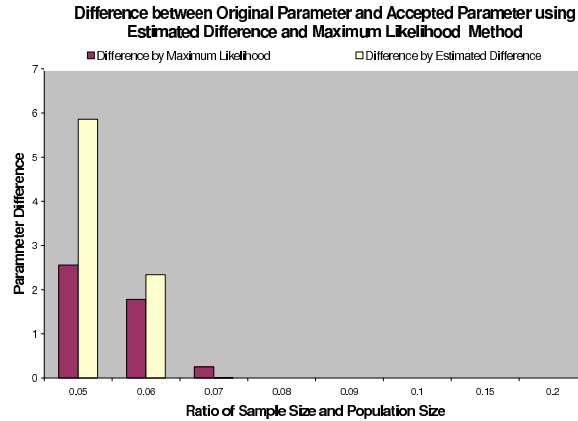


Figure 4.4: Difference between Original and Accepted Parameter using Estimated Difference and Maximum Likelihood

We impose the rule of probability distribution,

$$\sum_j P[\mu = \mu_j] = 1 \quad (4.14)$$

Now, a reading  $x$  is a Gaussian random variable with parameters  $\mu$  and  $\sigma$ . Let  $f(x|\mu)$  is the probability density function of  $x$  given a specified value of parameter  $\mu$ . Therefore, given a random sample  $\underline{x}$  of size  $n$ , the posterior probability distribution of the parameter  $\mu$  is given by:

$$P(\mu = \mu_j | \underline{x}) = \frac{f(\underline{x}|\mu_j)P[\mu = \mu_j]}{\sum_j f(\underline{x}|\mu_j)P[\mu = \mu_j]} \quad (4.15)$$

We define a risk function associated with any  $\mu_j$  as follows:

$$R(\mu, \mu_j | \underline{x}) = E[(\mu - \mu_j)^2 | \underline{x}] \quad (4.16)$$

This risk function calculates the amount of expected error if  $\mu_j$  is selected as an estimate of the true parameter. Clearly, the less the risk of error, the better the estimate. Therefore, the user will accept the  $\mu_j$  for which  $R(\mu, \mu_j | \underline{x})$  is minimum.

#### 4.10.2 Estimated Difference

The aggregated data is considered as weighted average of all sensor readings, the weight of a reading being an estimate of its correctness. Let  $\bar{x}$  denote the sample average. It is then

plausible to assume that the sample mean will be concentrated more around the correct average than any wrong average. Therefore, the server computes *error squares*  $E_j, j = 1, 2, \dots, k$  as defined below:

$$E_j = (\bar{x} - \mu_j)^2 \quad (4.17)$$

The server then accepts  $\mu_r$  if and only if  $E_r = \min_{1 \leq j \leq k} E_j$ .

### 4.10.3 Sample Size Determination With Known Population Standard Deviation

The procedure stated above requires a sample of fixed size  $n$ . It is obvious that the larger the sample size, the better the sample represents the set of all observations and consequently, the results are more accurate. But in reality, sampling in a sensor network is extremely expensive due to the stringent energy constraint and the expense is directly proportional with sample size. Therefore, the sample size has to be so determined that it involves minimal cost while helping to reach a decision as accurate as possible. Let us assume that each observation costs  $C$  units. Suppose the correct parameter value is  $\mu_0$ . Since we accept the  $\mu_j$  which is the closest to  $\bar{x}$ , the sample size must be so chosen that  $\bar{x}$  is a nearly accurate estimate of the original parameter value  $\mu_0$ . We define a loss function as  $E[\mu_0 - \bar{x}]^2$ . Thus, the total cost of involved in a sampling of size  $n$  is :

$$C(n, \mu) = E[\mu_0 - \bar{x}]^2 + An \quad (4.18)$$

To minimize this, the first derivative of the cost function with respect to  $n$  must be zero. Thus we need a solution of the following equation:

$$-\frac{\sigma^2}{n^2} + C = 0 \quad (4.19)$$

Thus, the optimal sample size is:

$$n_0 = \sigma/\sqrt{C} \quad (4.20)$$

### 4.10.4 Sequential Sampling Procedure with Unknown Population Variance

If the population variance  $\sigma$  is unknown, it is impossible to determine the sample size before sampling. An alternative procedure is sampling sequentially until we find a satisfactory outcome. Since  $\sigma$  is unknown, an ideal estimate for the sample size is  $\frac{s_n}{\sqrt{C}}$ , where  $s_n$  is an unbiased estimate of  $\sigma$ . The steps of the sequential procedure is given below:

- The user makes the aggregators to collect and send samples one at a request.
- $n$  is the number of samples drawn so far. Each time, the user calculates  $\bar{x}_n$  and  $s_n$  i.e., the sample mean and the sample standard deviation based on samples drawn so far.
- The user stops sampling request when for the first time for  $n \geq 2$  it gets  $n \geq \frac{s_n}{\sqrt{C}}$ .

According to Robbins [74], this rule terminates with probability 1.

#### 4.10.5 Maximum Likelihood Solution

We propose another deterministic approach for the general solution based on maximum likelihood estimation [73]. Let  $\Theta$  be the parameter space consisting of the parameters  $\mu$  and  $\sigma$ . Let  $f(x)$  be the pdf of the distribution of variable  $x$ . In this paper,  $f$  is a Gaussian pdf with parameter space  $\Theta = \{\mu, \sigma\}$ . The likelihood function of parameter  $\mu$  corresponding to sample values  $x_1, x_2, \dots, x_n$ , denoted by  $L(\mu|x_1, x_2, \dots, x_n)$ , is defined as:

$$L(\mu|x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{\Theta}(x_i) \quad (4.21)$$

$L(\mu|x_1, x_2, \dots, x_n)$  can be regarded as the probability that given the observations  $x_1, x_2, \dots, x_n$ ,  $\mu$  is the true parameter. Note that since  $\sigma$  is unknown, we compute  $f_{\Theta}(x_i)$  using an estimate of  $\sigma$ . The estimate of  $\sigma$  is given as:

$$\hat{\sigma} = \sum_{i=1}^n (x_i - \bar{x}) / (n - 1) \quad (4.22)$$

where  $\bar{x} = \sum_{i=1}^n x_i / n$ . The estimated pdf is denoted as  $\hat{f}_{\Theta}(x_i)$  and the estimated likelihood function is denoted as  $\hat{L}(\mu|x_1, x_2, \dots, x_n)$ .

Therefore, given the set of distinct averages  $\mu_1, \mu_2, \dots, \mu_k$  from different aggregators and the sample values  $x_1, x_2, \dots, x_n$ , the server computes  $\hat{L}(\mu_i|x_1, x_2, \dots, x_n)$  for  $i = 1, 2, \dots, k$  and accepts  $\mu_r$  if

$$\hat{L}(\mu_r|x_1, x_2, \dots, x_n) = \max_{1 \leq i \leq k} L(\mu_i|x_1, x_2, \dots, x_n) \quad (4.23)$$

This method is deterministic and easy to compute.

#### 4.10.6 Comparative Performance Evaluation of Estimated Difference and Maximum Likelihood Method

We have studied and compared performances of the two methods described above using real temperature data from a region of Louisiana collected at a particular time and date. The range of temperature varies from 50 degree to 75 degree. In the simulation setup, we have deployed 1000 sensors throughout the whole region. There are 5 aggregator nodes who collect data individually from all the sensors and compute the average temperature. We have selected the number of faulty aggregators at random. Those faulty aggregators inject a random error varying within  $(-10, 10)$  into the sensor readings. The aggregators compute the temperature average and send to the user. The user then uses Estimated Difference and Maximum Likelihood Method to determine which average value is to accept. We study the results by varying sample size. We notice that both the methods give accurate solution with a sample size as small as 8 percent of the total population size. However, according to the simulation, Estimated Difference method outperforms Maximum Likelihood method.

#### 4.10.7 Sequential Probability Ratio Test

Testing of statistical hypothesis is a well-known procedure of determining the validity of an assertion about an unknown parameter value. There are many standard tests in the literature of statistics for testing different kinds of hypothesis. However, all of them involve a fixed sample size. Fixed sample size tests introduce the problem of further determining an optimal sample size. In this paper, we consider a Normal population with unknown mean and variance. The variance being unknown, it is impossible to determine the optimal sample size. For stringently energy constrained sensor networks, however, sampling is extremely expensive. Therefore, in this paper, we consider Sequential Probability Ratio Test (SPRT) [73] which minimizes the sampling cost by using sequential sampling procedure. SPRT test can only be applied when there are two possible values of a parameter and the task is to assert one value against other based on random samples. The test is described below:

SPRT test is performed to test a hypothesis  $H_0 : \mu = \mu_0$  against an alternative  $H_1 : \mu = \mu_1$ . Let  $f_{0n}$  and  $f_{1n}$  denote the joint pdf's of sample values  $x_1, x_2, \dots, x_n$  i.e., the vector  $\underline{x}$  under  $H_0$  and  $H_1$  respectively.  $f_{in} = \prod_{j=1}^n f_{\mu_i}(x_j)$ ,  $i = 0, 1$ . We define  $\lambda_n(\underline{x}) = \frac{f_{1n}(\underline{x})}{f_{0n}(\underline{x})}$ . The SPRT is defined as a rule that states:

1. if at any stage of sampling,  $\lambda_n(\underline{x}) \geq A$ , stop and reject  $H_0$  and accept  $H_1$ ;
2. if at any stage of sampling,  $\lambda_n(\underline{x}) \leq B$ , stop and accept  $H_0$ ;
3. if  $B < \lambda_n(\underline{x}) < A$ , continue sampling by taking another observation  $x_{n+1}$ .

Here  $A$  and  $B$  are constants which are determined so that the test will have strength  $(\alpha, \beta)$ ,  $\alpha$  being the probability that a true hypothesis will be rejected i.e.,  $P(\text{reject } H_0 | H_0)$  and  $\beta$  being the probability that a false hypothesis is accepted i.e.,  $P(\text{accept } H_0 | H_1)$ .  $\alpha$  and  $\beta$  are predetermined constants on which the value of  $A$  and  $B$  depends,  $0 < \alpha, \beta < 1$ . The user should select small values of  $\alpha$  and  $\beta$  in order to keep the error probabilities small. In other words, small values of  $\alpha$  and  $\beta$  enables the user to accept a correct result with high probability. We state two significantly important theorems proposed by Wald [73] which characterize the SPRT test.

**Theorem 1** *For the SPRT with stopping bounds  $A$  and  $B$ ,  $A.B$ , and strengths  $\alpha$  and  $\beta$ ,  $A = \frac{1-\beta}{\alpha}$ ,  $B = \frac{\beta}{1-\alpha}$ .  $0 < \alpha < 1$ ,  $0 < \beta < 1$ .*

**Theorem 2** *The SPRT terminates with probability 1 under both  $H_0$  and  $H_1$ .*

## 4.11 Conclusion

In this paper, we address the problem of stealthy attack in a sensor network which results in wrong sensor readings. We propose a probabilistic accuracy model based on which an aggregator can perform weighted data aggregation in order to make the aggregated data close to the true value. We introduce a multi-aggregator setup which enables a user receive aggregated data for the same region from multiple aggregators and detect any error by comparing those results. Several statistical testing methods are proposed using which the user can determine the correct information value with high probability.

# Chapter 5

## Energy Optimized Routing for Passive Security

### 5.1 Introduction

Wireless sensor networks are significantly vulnerable to data attacks by adversaries due to large scale, autonomous operation and data flow over insecure wireless channels. Preventive security measures can be broadly classified into two categories - active security mechanisms and passive security mechanisms. We have provided some novel active security mechanisms in the previous chapters. In this chapter, we deal with the problem of passive security. Stringent energy constraints in a sensor network result in expedited network partition, thereby enabling an adversary to perform powerful attacks. Many security protocols for sensor networks [69, 77] assume that the number of nodes that an adversary can compromise cannot exceed certain number. Therefore, the smaller the network size, the more vulnerable it is to enemy attack. Uneven energy distribution across a sensor network leads to quick network partition, thereby reducing the size of the network component connected to the base station and exposing it to the risk of powerful attack. Therefore, imposing uniform energy consumption across the network significantly increases robustness of many security protocols, thus contributing passively to the safety of the network. Among the various upper-level network functions, routing in particular must be extremely energy-efficient as improperly chosen routes could lead to uneven residual energy distribution across sensors and expedite network partition. Therefore, the objective of our passive security mechanism is to provide an energy efficient routing protocol which increases network lifetime by inducing uniform energy consumptions across the network. Since each sensor may be participating simultaneously in several routes

(via multiple data-aggregation trees to the sink(s)) [19], the energy consumption of a sensor node is inversely proportional to the number of routes it participates in. Thus routes must be chosen carefully according to the following general principles: First, data should be routed over paths in which participating nodes have higher energy levels relative to other non-participating nodes. Second, shorter paths should be favored since they involve fewer participating sensors thus reducing overall energy consumption. Third, routing protocols should be as localized (distributed) as possible to reduce the scalability and robustness limitations of centralized top-down solutions.

For sensor networks operating under stringent energy constraints, strictly localized (fully distributed) routing algorithms which exploit only local network state information are more energy-efficient than those requiring full network state information. While it is true that such protocols will have better scalability (due to the ability of nodes to act independently) and robustness (resiliency to centralized failures) properties, there may be situations in which *limited* global information/propagation will be beneficial to the network. Intuitively, strictly localized routing protocols will take more time to adapt *efficiently* to certain boundary conditions (as we describe in the example below). Thus a challenging issue is to find ways of exploiting the benefits of collecting global network state information (flow rates, energies etc.) versus the overhead and other limitations of doing so.

We note that the routing choices of sensor nodes under the above constraints are a natural fit for a decision-making framework. Therefore we model the problem of finding energy-efficient routing paths with bounded lengths using a decision-theoretic paradigm in which routes are chosen using *profit* functions calculated at each sensor. The profit function is defined in such a way that it induces each node to link to the healthiest possible node while forming short paths. These two factors ensure that only a small number of relatively healthy nodes participate in routing, thereby reducing overall energy consumption and potentially delaying network partition. We define *length-energy-constrained* (LEC) optimal path as a path in which each sensor selects its best possible strategy given the choices of all other sensors. While computing this path is NP-hard in arbitrary sensor networks, we show that it can be found in polynomial time (in a distributed manner) in sensor networks operating under a geographic routing regime. However, the distributed algorithm requires a lot of state information to be stored in each sensor. In this paper, we propose a **nearly stateless**, scalable and easily implementable distributed protocol which approximately calculates an LEC optimal path. We also propose two other protocols to calculate optimal paths under some variations of the proposed profit function.

We classify our protocols as *quasi*-strictly-localized. While sensors mostly obtain information only from a small set of immediate neighbors, limited state information is periodically propagated through a restricted neighborhood of the network. We show in this paper that using this limited global information on node-energies combined with local geographic for-

warding can remarkably improve network lifetime. In a sense, the proposed protocols are similar in spirit to GEAR [86] which is a localized but not strictly localized protocol since it requires nodes to know their distance (hops) from the destination.

The first protocol *length-energy constrained geographic routing* (LCGR) uses a distributed algorithm to determine optimal paths defined by the above routing formation procedure in a geographically routed sensor network. The second protocol *max-min energy-constrained geographic routing* (MEGR) finds optimal paths for a simplified "team version" of the route formation process while the third protocol *threshold-energy constrained geographic routing* (TCGR) uses a threshold constrained heuristic for finding optimal paths. All five protocols use *reverse directional flooding* to find optimal routes by combining limited global information with local geographic forwarding. While reverse directional flooding involves some extra overhead and limited node energy, global energy information obtained through this process ensures a significant tradeoff in terms of balancing energy across the network and network lifetime.

The key features that distinguish our protocols from other related routing protocols are summarized below.

- Network lifetime maximization protocols such as [82] require global information on current data/packet flow rates from each sensor to sink(s). Linear programming and other techniques are then used to calculate routes for maximizing network lifetime. These protocols require significant global network state information and are consequently difficult to implement. In contrast, the proposed protocols require very limited network state information to calculate threshold values.
- Since energy is a critical resource in sensor networks, depleted regions, i.e regions with low residual node energies must be detected and bypassed by routing paths as quickly as possible. This is analogous to congestion in wired networks. We propose a new technique for indicating the onset of energy depletion in regions by using *energy depletion indicators*. We ensure that all five protocols provide energy-balanced routing by using this indicators in conjunction with *threshold energy levels*. In fact, the minimum energy level of any node in TCGR is continuously bounded below by at most one packet transmission cost from the threshold value.
- Geographic sensor network routing algorithms such as GPSR [83] and GEAR [86] are well-known protocols which combine energy aware neighbor selection with geographic forwarding. The neighbor selection procedure in GEAR is based on a parametric combination of local information such as node energy consumed to date, and node distance to destination. While this is a very elegant and easily implementable protocol, there are situations in which the protocol will be slow to adapt to changing energy distributions,

due to its predominantly localized nature. For example, consider a region in a sensor-net which is intersected by multiple routes. Nodes in this region will tend to deplete energy at higher rates. While GEAR will take a considerable amount of time to avoid this region through localized rerouting, our proposed protocols using a limited global threshold mechanism, will be able to detect higher energy depletion rates quickly and find a new route in comparatively lesser time.

- A potential drawback of protocols such as [86], [83], [88] which use local geographic forwarding is that significant backtracking is required when a hole is encountered. This situation is completely eliminated by our first two protocols since packets are forwarded according to the periodically updated routing table residing at each node.
- In protocols such as [86] [83] [81] [80] [87], a single routing path (typically, the least energy path) is utilized continuously until a node's energy is completely exhausted. While the motivation behind this approach is to save energy consumption at individual sensor nodes, this might lead to unintended consequences such as the expedited partition of the network. Our protocols overcome this drawback by selecting new length-energy-constrained routing paths periodically. ([84] also does this but in a probabilistic manner for non least-energy cost paths).

We have evaluated our routing protocols using the ns-2 simulator. Simulation results indicate that all the protocols are enormously effective in reducing energy deviation, thereby leading to equitable residual energy distribution across the sensor network. Thus the protocols should have a significant impact on sensor network survivability.

## 5.2 Related Works

There is a large body of existing research on energy-efficient routing protocols for sensor networks. While most research ([90], [91], [42]) focuses on the energy cost of routing paths as the critical metric, several recent works consider the issue of routing path lengths in the context of localization as well. For probabilistic routing protocol where non least-energy cost paths are chosen periodically to help in energy balancing. Geographic routing, as described in GPSR [83] and GEAR [86] is a popular technique for reducing the length of routing paths by restricting the forwarding neighborhood choices of sensors according to geographic direction. GEAR [86] is a seminal localized routing protocol in which a node attempts to balance energy across all its neighbors while finding geographically oriented shortest paths to the sink. In [87], routing has been considered under constraints of packet reception/packet loss and distance. K. Seada et al. in [87] proposed several packet forwarding

metrics which outperform geographic forwarding. However, there is no unified analytical model in the literature that explicitly considers routing under both the constraints of energy efficiency and path length. In protocols such as [86] [83] [81] [80] [87], a single routing path (typically, the least energy path) is utilized continuously until a node's energy is completely exhausted. [84] overcome this drawback by selecting new length-energy-constrained routing paths periodically in a probabilistic manner for non least-energy cost paths.

### 5.3 Analytical Model of Length-Energy-Constrained Routing

With current technology, communication energy costs typically outweigh processing and sensing costs in sensor networks. Thus the longevity of a sensor node depends heavily on the number of routing paths it participates in. Improperly chosen routing paths will lead to uneven energy consumption across sensors; highly non-uniform residual node energy might also expedite network partition. Therefore routing protocols between any source and destination pair must be designed to dissipate energy equitably over sensors.

One possible approach is to prevent low energy nodes from taking part in a route as long as they are energy-deficient relative to their neighbors. However, a route that focuses only on energy efficiency may be undesirably long (in terms of hop count) since the lowest energy-cost path need not be the shortest. Longer paths will result in energy depletion at more sensors while also increasing delay. While there are several existing protocols in the literature that focus exclusively on either of these issues, there is no unified analytical model that explicitly considers routing under both the constraints of energy efficiency and path length.

In this paper, we model sensors as intelligent agents and propose a decision theoretic paradigm for solving the problem of finding energy-optimal routing paths with bounded path length. We propose a framework in which each sensor receives a credit for making a link to a neighbor. However, it has to pay certain cost for participating in a path and contributing to the length of the path thus formed. While the credit is a function of residual energy of neighboring nodes, the cost is a function of the path length. Thus, each sensor makes a profit (net credit) by making a link to a neighbor. A route is formed as a consequence of decisions taken by all the sensors. Intuitively, the optimal route should produce the maximum possible profit for each individual sensor, given the choices of all other sensors. In other words, given the choices of all other nodes, no node will be better-off by deviating from the current choice in an optimal path. We now formally define our analytical model.

Let  $S = \{s_1, s_2, \dots, s_n\}$  be the set of sensors in the sensor network participating in the routing formation procedure. Let  $L_1$  and  $L_2$  be a pair of source and destination nodes using sensors in  $S$  as intermediaries<sup>1</sup>. Data packets are to be routed from  $L_1$  to  $L_2$  through an optimally chosen set  $S' \subset S$  of intermediate nodes by forming communication links. Note that we do not consider multicast communication between sets of source and destination nodes in this paper.

**Strategies:** Each node has a strategy for making links. A strategy is a binary vector  $l_i = (l_{i1}, l_{i2}, \dots, l_{ii-1}, l_{ii+1}, \dots, l_{in})$ , where  $l_{ij} = 1$  ( $l_{ij} = 0$ ) represents sensor  $s_i$ 's choice of sending/not sending a data packet to sensor  $s_j$ . Since a sensor typically relays a received data packet to only one neighbor, we assume that a node forms only one link for a given source and destination pair of nodes. In general, a sensor node can be modeled as having a mixed strategy [79], i.e., the  $l_{ij}$ 's are chosen from some probability distribution. However, in this paper we restrict the strategy space of sensors to only pure strategies. Furthermore, in order to eliminate some trivial optimal paths, each sensor's strategy is non-empty and strategies resulting in a node linking to its ancestors (i.e. routing loops) are disallowed. Consequently, the strategy space of each sensor  $s_i$  is such that Prob.  $[l_{ij} = 1] = 1$  for exactly one sensor  $s_j$  and Prob.  $[l_{ij} = 1] = 0$  for all other sensors, such that no routing loops are formed.

**Profit Function:** Let  $l = l_1 \times l_2 \times \dots \times l_n$  be a strategy resulting in a route  $\mathcal{P}$  from source to destination node. Each sensor on  $\mathcal{P}$  receives a profit from participating in this route. The profit of a sensor  $s_i$  which links to node  $s_j$  in  $\mathcal{P}$  is then defined as:

$$\pi_i(l) = E_j - \xi L(\mathcal{P}) \quad (5.1)$$

where  $E_j$  is the residual energy level of node  $s_j$  and  $L(\mathcal{P})$  the length of routing path  $\mathcal{P}$ .  $E_j$  represents a credit earned by  $s_i$ , thus inducing it to forward data packets to higher energy neighbors. The parameter  $\xi$  represents the proportion of path length costs that are borne by sensor  $s_i$ . Choosing  $\xi$  as a positive constant or proportional to path length will inhibit the formation of longer routing paths. Conversely, setting  $\xi$  zero or inversely proportional to path lengths will favor the formation of paths through high-energy nodes. We choose  $\xi$  as a non-zero positive constant for this route formation procedure. Thus each sensor will forward packets to its maximal energy neighbor in such a way that the length of the path formed is bounded.

A *length-energy-constrained* optimal path is defined as the path in which all participating sensors have chosen their best-response strategy, i.e., the one that yields the highest possible credit given the strategies of other nodes. This is the optimal Length Energy-Constrained (LEC) route in the sensor network for the given source and destination pair. Note that the

---

<sup>1</sup>In general, sensors in  $S$  will be simultaneously participating in routing paths between several such pairs.

process of determining the LEC route requires each node to determine the optimal paths formed by each of its possible successors on receiving its data. The node then selects as next neighbor that node, the optimal path through which incurs the highest profit.

We present another alternative payoff model which also encapsulates the path-length constraint in a different format. In this model, each node pays a proportion of the total energy consumption along the selected path as a cost. Let  $l = l_1 \times l_2 \times \dots \times l_n$  be a strategy resulting in a route  $\mathcal{P}$  from source to destination node. Let  $E(\mathcal{P})$  be the total energy consumption for this path. Therefore, the profit function is given as:

$$\pi_i(l) = E_j - \eta E(\mathcal{P}) \quad (5.2)$$

Similar to the previous model,  $\eta$  is the proportion of total energy consumption cost borne by sensor  $s_i$  and has similar properties. Note that since total energy consumption is proportional to the path length, this model too encapsulates the process of decentralized route formation by making sensor nodes cooperate to achieve a joint goal (shorter routing paths) while optimizing their individual benefits.

Similar to an LEC optimal route, a *total-energy-constrained* (TEC) optimal path is the one in which all participating sensors have chosen their best-response strategy. All the characteristics of a TEC optimal path are the same as that of an LEC optimal route.

**Theorem 3** *Let  $\hat{\mathcal{P}}$  be the optimal LEC route for a pair of source and destination nodes in an arbitrary sensor network. Computing  $\hat{\mathcal{P}}$  is NP-Hard. Similarly, computing a TEC optimal path is also NP-hard.*

**Proof:** We will prove this theorem for an LEC route by reduction from Hamiltonian path. Our proof relies on constructing a specific example of a network with a particular source and destination pair, in which an optimal path contains a Hamiltonian path as subpath.

Consider an arbitrary graph  $G_1 = (V_1, E_1)$  where  $V_1 = v_1, v_2, \dots, v_n$ . We construct a sensor network  $G = (V, E)$  as shown in Figure 5.1 with the following parameters: The vertex set  $V$  is the union of vertex set  $V_1$  and nodes  $S, A$  and  $D$ ,  $S$  and  $D$  being the source and the destination correspondingly. The edge set  $E$  is the union of disjoint edge sets  $E_1, E_2$  and  $(A, D)$  where  $E_2 = (S, v_i) \cup (v_i, A)$  for all  $i$ . Let the energies of all the nodes in  $G_1$  be  $n + e$  and that of node  $A$  be  $e$ , where  $e \geq 1$ . In this case, after getting a packet from  $S$  any node in  $G_1$  has two choices: a) forward the packet to  $A$ , b) forward the packet to any node in  $G_1$ . Let  $\xi = 1$  for sake of simplicity. In case a), the profit of this node is  $e - 3$ . In case b), the maximum path length possible is  $n + 2$  if there exists a Hamiltonian path in  $G_1$ . Hence, the minimum possible profit of each node in  $G_1$  is  $n + e - (n + 2)$  which is greater than  $e - 3$ . Therefore, the optimal path from  $S$  to  $D$  consists of a Hamiltonian path in  $G_1$  if there exists any.

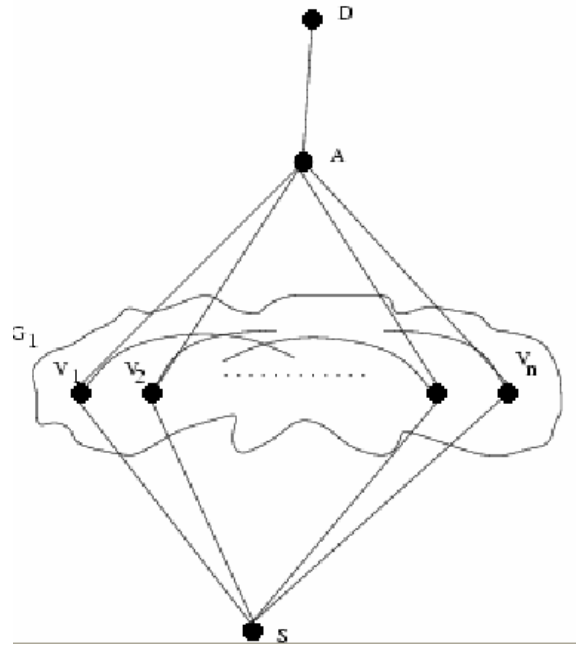


Figure 5.1: An optimal LEC path is NP-hard.

With a similar reasoning, we can prove that computing a TEC optimal path is also NP-hard. ■

**Theorem 4** *Let  $S$  be any sensor network in which sensor are restricted to following a **geographic routing** regime. In other words, the strategy space of each sensor includes only those neighbors geographically nearer to the destination than itself. Then  $\hat{\mathcal{P}}$  can be computed in polynomial time in a distributed manner. The result holds for a TEC optimal path as well.*

**Proof:** Due to space limitations, we only provide an outline of the proof. Section 4.2 describes our protocol for calculating the optimal path  $\hat{\mathcal{P}}$  which can be obtained in  $O(N+E)$  steps for an arbitrary  $N$ -node geographically routed sensor network with  $E$  edges. The result hinges on the observation that in a geographically routed network, the intersection of all feasible paths from the source to a node  $s_i$  and from  $s_i$  to the destination node is exactly  $s_i$ . Thus one can compute the union of optimal paths from source to sensor and sensor to destination. ■

Next, we identify sufficient conditions under which the optimal LEC path coincides with other commonly used routing paths. For brevity, we state these results without proof.

**Proposition 9** Let  $E_{max}^i$  and  $E_{min}^i$  denote the maximum and minimum neighbor node energies at sensor  $s_i$ . Then the shortest path from  $L_1$  to  $L_2$  will be optimal if

$$E_{max}^i - E_{min}^i < \xi(\delta S) \quad (5.3)$$

holds at each sensor node  $s_i$  on the shortest path, where  $(\delta S)$  is the difference between the shortest and second shortest paths from  $s_i$  to  $L_2$ .

**Proposition 10** Let the maximal-energy neighbor path denote the one obtained by following the maximal energy nodes from  $L_1$  to  $L_2$  such that a path is formed. Then we have, The maximum energy neighbor path will be optimal if

$$(\delta E) > \xi(l_h - l_s) \quad (5.4)$$

holds at every node  $s_i$  on the path, where  $(\delta E)$  is the difference in energies between the maximal and second maximal neighbors of  $s_i$  and,  $l_h$  and  $l_s$  are the lengths of the maximal energy and shortest paths from  $s_i$ , respectively.

## 5.4 Distributed Protocol Implementation

In the previous section, we have mentioned that in a geographic routing regime, an LEC optimal path can be computed in polynomial times in a distributed manner. However, such an algorithm will require a lot of state information to be stored in each sensor node, thereby increasing overall energy consumption. In this section, we propose and describe three different *nearly stateless and scalable* protocols for finding an energy-efficient route and balancing energy across the sensor network as a concomitant side-effect. This section is divided into three subsections, each subsection describing a protocol.

### 5.4.1 Length-energy Constrained Geographic Routing Protocol (LCGR)

This protocol is a distributed implementation of the optimal LEC route formation process. In this protocol, each node calculates its highest profit according to the model described above and forms the optimal path. Since node energy levels are changing continuously in a sensor network due to sensing, processing and routing operations, both the optimal path and the threshold need to be recomputed periodically. Thus the proposed protocol operates in two different phases: data transmission and path determination. During the path determination phase, the calculation of the optimal path and the threshold value takes place. The protocol is described below in details:

**Data Transmission Phase:** During this phase, data packets are transmitted from a source node  $L_1$  to a destination node  $L_2$  through the optimal path (with least energy weakness). Each data packet also potentially collects information about the energy consumption en route, by keeping track of residual energy levels of nodes on the path. When energy levels of a given critical number of nodes fall below a certain threshold, the data transmission phase ends and the new optimal path determination phase begins.

The fundamental steps of the data transmission phase are as follows:

- Each data packet is marked by the source node with the geographical position of the destination node and with a threshold value  $th$ . Each data packet contains a special  $n$ -bit Energy Depletion Indicator (EDI) field, where  $n \ll$  packet size.
- Each sensor node receiving a data packet determines whether its energy level has fallen below the threshold  $th$ . If so, and the EDI field in the data packet is not exhausted, the node sets a single bit in the EDI field. Then it forwards the packet to the best next-hop neighbor according to its routing table. We assume that before the network starts any activity, all ordinary sensor nodes have the same energy level. Therefore, during the first data transmission phase, the best next-hop neighbor of a node is the one which is geographically nearest to the destination node. In all other phases, the routing table is updated according to the optimal LEC path calculation.
- If the destination node gets a data packet with all  $n$  bits in the EDI field set to 1, it triggers a new optimal path selection procedure. Note that the length of the EDI field is an empirical value that must be chosen carefully, as discussed in section 5.2.

**Calculation of the Threshold Value:** The threshold value  $th$  plays a very important role in the data transmission phase since it is used to provide an approximate indication that the current optimal path has become obsolete. Intuitively,  $th$  must be a function of the current residual node energy levels in the network. In this paper, we use the following function for all three protocols:

$$th = \beta E_{min} \quad (5.5)$$

where  $0 < \beta < 1$  and  $E_{min}$  is the maximum of minimum node energy levels on all geographic routing paths to the destination  $L_2$ . Since  $E_{min}$  changes with time, the threshold is recalculated in each path determination phase, consistent with the current energy distribution across the network.

**Path Determination Phase:** This phase begins when the destination node receives critical EDI information and ends when the source node has updated its routing table and recalculated the threshold value. The principle steps are as follows:

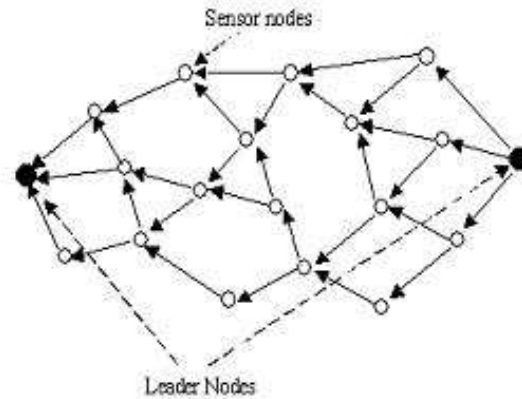


Figure 5.2: Reverse Directional Flooding

- The destination node  $L_2$  triggers this phase by flooding the network with control packets along the geographic direction of the source node  $L_1$  (Figure 5.2). Note that this *reverse directional flooding* occurs in the direction opposite to that of data transfer.
- Each node forwards *exactly one* control packet to all its neighbors in the geographic direction of  $L_1$ . Each control packet contains three fields: the given node's residual energy level, a length field  $L(P)$  that indicates the length of the current *optimal* partial path from that node to  $L_2$  and a max-min energy field  $EM_p$  that indicates the maximum of the minimum node energy levels on all partial paths to  $L_2$  originating at the given node.  $L(P)$  is calculated in the iterative way described below.
- On receiving the first control packet, each node sets a timer for a prefixed interval  $T$ . This time-period should be large enough for the node to receive future control packets from all or most of its upstream neighbors (corresponding to different partial paths from the upstream nodes to  $L_2$ ), but not so large as to cause high delays. With each arriving control packet, the node calculates, updates and stores the highest  $E - \xi L(P)$  value seen so far, where  $E$  and  $L(P)$  are the residual energy level and optimal partial path length to  $L_2$  from that upstream neighbor. It also updates and stores the highest  $EM_p$  value seen so far. *However, if its own energy level  $E_i$  is lower than all these  $EM_p$  values, it stores  $E_i$ .* With each control packet, the given node also updates its routing table for destination  $L_2$  to point to the node from which it will receive the highest profit. Note that the choice of this optimal neighbor is independent of partial path

lengths from  $L_1$  to the given node and is in fact the upstream node with the highest  $E - \xi L(P)$  value .

- When its timer expires, this node creates a control packet with  $L(P)$  field set as the length of the current optimal partial path to  $L_2$  (via the highest profit neighbor). The control packet also contains the current  $EM_p$  and residual energy fields and is forwarded to all its neighbors in the geographic direction of  $L_1$ . Control packets arriving after the timer expires are discarded.
- Eventually,  $L_1$  begins receiving control packets and sets its timer. Its value of  $T$  can be determined in many ways depending on the specific requirements of applications. In this paper, we calculate  $T$  to ensure that most of the paths from  $L_1$  to  $L_2$  are included in the optimality calculations. If ( $D_{max}$ ) is the maximum transmission delay between two nodes, the value of  $T$  is determined as ( $MINHOP * D_{max}$ ), where  $MINHOP$  is an estimate of the shortest path from  $L_1$  to  $L_2$ . This value can be estimated a priori using GPSR routing [83] before the first data transmission phase. Note that the given value of  $T$  allows control packets from paths up to twice the length of the shortest path to be forwarded to  $L_1$ . Also note that  $D_{max}$  is a function of the specific MAC-layer protocol being implemented in the sensor network. Finally, when the timer expires at  $L_1$ , it sets its routing table and calculates the new threshold value  $th$  using  $E_{min}$  as the highest received  $EM_p$  value. The next data transmission phase can now begin.

### 5.4.2 Max-min Energy-constrained Geographic Routing Protocol (MEGR)

For comparative purposes with LCGR, we consider an alternative protocol implementing a simplified ‘team’ version of the original LEC route formation procedure. MEGR has the same overhead as LCGR but computes optimal paths using the following team path heuristic: each node on a path shares the profit of the worst-off node on it. Formally, let  $\mathcal{L}$  be the set of all distinct paths from a particular source and destination pair. Let  $E_{min}(\mathcal{P})$  be the smallest residual energy value on path  $\mathcal{P}$ . Then the max-min optimal path is defined as:

$$\hat{P} = \operatorname{argmax}_{\mathcal{P} \in \mathcal{L}} (E_{min}(\mathcal{P}) - \xi |\mathcal{P}|) \quad (5.6)$$

For simplicity, we set  $\xi$  to zero. However, the protocol can be easily implemented for non-zero values of  $\xi$ . We interpret the optimal path under this condition as follows: Given any path  $\mathcal{P}$ , the durability of the path is inversely proportional to  $E_{min}(\mathcal{P})$ . A path with lower average energy but higher minimum energy should last longer than a route with the opposite attributes since the least energy node is the first to terminate and make that route obsolete.

The inverse of the minimum node energy on a given path reflects the *energy weakness* of the path. Thus MEGR will select an optimal path with the least energy weakness. The protocol is implemented in the same manner as LCGR with data transmission and path determination phases.

### 5.4.3 Min-range Energy-constrained Geographic Routing Protocol (REGR)

We present an alternative path optimality criteria which defines an optimal path as the one on which the energy difference of the maximum energy node and the minimum energy node is the minimum. Formally, let  $\mathcal{L}$  be the set of all distinct paths from a particular source and destination pair. Let  $E_{min}(\mathcal{P})$  and  $E_{max}(\mathcal{P})$  be the smallest and the highest residual energy value on path  $\mathcal{P}$  respectively. Then the min-range optimal path is defined as:

$$\hat{P} = \operatorname{argmin}_{\mathcal{P} \in \mathcal{L}} (E_{max}(\mathcal{P}) - E_{min}(\mathcal{P})) \quad (5.7)$$

Note that a min-range optimal path too has the objective of keeping the residual energy distribution uniform. This heuristic can be implemented in the same way as MEGR.

### 5.4.4 Threshold-energy Constrained Geographic Routing Protocol (TCGR)

In the two protocols described above, an optimal route is found in the path determination phase. This route is fixed until another path determination phase is triggered by the energy depletion indicator. However, it may happen that only one (or very few) node(s) in a route become critically energy-deficient. A new path selection procedure will not be initiated until all bits of the energy depletion indicator are set to one which is unfair to the given node(s). To overcome this drawback, we propose another protocol with the same max-min energy metric controlled by two threshold-energy values. TCGR differs from the other two protocols in that two threshold values are calculated during the threshold determination phase while the optimal path is determined during the data transmission phase. The key steps of TCGR are described below:

Data Transmission Phase:

- Each data packet is marked by the source node with the geographical position of the destination node and with two threshold values  $min_{th}$  and  $max_{th}$ . These two threshold

values are defined as:

$$min_{th} = \alpha E_{min}, \quad 0 < \alpha < 1 \quad (5.8)$$

$$max_{th} = \rho min_{th}, \quad 1 < \rho < \frac{1}{\alpha} \quad (5.9)$$

Note that  $E_{min}$  is the max-min residual energy of all geographic paths from the source to the destination.  $\alpha$  is an ‘inverse’ density parameter that impacts the the set of feasible threshold bounded routes to the destination.

- Each sensor node receiving a data packet forwards it to that neighbor with energy level higher than  $min_{th}$  which is geographically closest to the destination.
- If all neighbors that are closer to the destination have energy level below  $min_{th}$ , i.e., when there is a hole, a node selects a neighbor whose energy level is highest above  $min_{th}$ . When there is no such neighbor, the node sends back the packet to its predecessor with a special message that the path is blocked. The predecessor node updates its routing table by placing the next best geographic neighbor as the next-hop and forwards the data packet to this neighbor. Note that there is a certain amount of backtracking involved which contributes to energy inefficiency. This is because we always to attempt to select the shortest path first. However by appropriate choice of parameters as described in section 5, the amount of backtracking can be reduced.
- If the source receives a blocked data packet, this implies that a new threshold value  $min_{th}$  must be computed, as described in the next section.
- Each sensor node receiving a data packet determines whether its energy level is below another threshold  $max_{th}$ , a function of  $min_{th}$ . If the energy level of the node is below  $max_{th}$  and if the  $n$  bit EDI field in the data packet is not exhausted, the node sets a single bit in the EDI field.
- If the receiver node gets a data packet with all  $n$  bits in the EDI field set to 1, it triggers a new selection procedure for  $min_{th}$ .
- Alternately, if the destination receives  $m$  data packets with at least one bit in the EDI field set to 1 a new selection procedure for  $min_{th}$  is triggered. Note that a single node on the current path with energy level below  $max_{th}$  continues to remain on the path until its energy level falls below  $min_{th}$  ( At this point packets are rerouted around this node ). We want new thresholds to be computed when multiple nodes *sequentially* fall below  $min_{th}$ ( Note that this scenario is not captured by the previous protocols). Let

$k \ll m$  be the approximate number of transmitted packets required for a node's energy to fall from  $max_{th}$  to  $min_{th}$ . Choosing  $m = kn$  will ensure that new thresholds will be computed in case of sequential energy depletions.

Note that the protocol as described above bounds the minimum energy level of any node from below by at most one packet transmission cost from  $min_{th}$ . As soon as a node's energy falls below this value it is cut off from participating until a new threshold is computed.

Threshold Determination Phase: During this phase, the current value of  $E_{min}$  is calculated using a procedure similar to the one described in the *path determination* phase of the first two protocols.

## 5.5 Selection of Parameters

There are two main issues related to practical implementation efficiency of our protocol.

- To avoid unnecessary energy expenditures, control packets must be prevented from hanging around the network after the path determination phase. The parameter MIN-HOP serves this purpose.
- There is a tradeoff between energy consumption involved in flooding vs the gain in network lifetime due to equitable energy distribution among sensors. Therefore, frequency of invocation of the path determination phase is an important parameter. We experimentally model this using  $\beta$ , EDI,  $\alpha$  and  $\rho$  as described later.

### 5.5.1 Selection of Energy Depletion Indicator

The length of the EDI field determines the maximum number of critical nodes allowed during a data transmission period, thus regulating the duration of this phase. This parameter is empirical and can be modified by the source at the beginning of each data transmission phase using the following rule:

$$EDI_{current} = \gamma EDI_{prev1} + (1 - \gamma) EDI_{prev2} \quad (5.10)$$

where  $0 \leq \gamma \leq 1$  and  $EDI_{prev1}$  and  $EDI_{prev2}$  are the previous values of the EDI field.  $\gamma$  should be chosen according to the specific requirement i.e. whether the duration of the data transmission phase should be increased or decreased and to what extent.

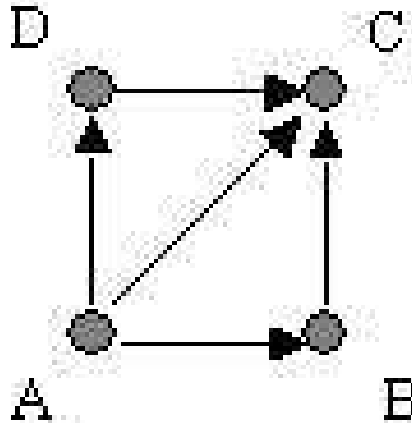


Figure 5.3: Possible Control Packet Optimization

### 5.5.2 Selection of $\beta$ , $\alpha$ and $\rho$

In LCGR and MEGR, data transmission ends when residual energy levels of at least  $n$  nodes on the current path fall below threshold  $th$ . The smaller the value of  $\beta$ , the larger the useful data transmission phase. In TCGR, the value of  $min_{th}$  is proportional to  $\alpha$ . Lower values of  $\alpha$  will increase the number of feasible paths from source to destination nodes. However, this will also undesirably increase the number of energy deficient nodes participating in routes. Similarly  $\rho$  and the length of the EDI field control the duration of data transmission, since reverse directional flooding starts when the residual energy levels of at least EDI number of nodes participating in the path fall below threshold  $max_{th}$ . Therefore these empirical parameters should be carefully determined using observed energy depletion rates and traffic patterns.

### 5.5.3 Overhead Due to Reverse Flooding

The proposed protocol uses reverse directional flooding to determine a new optimal path. The advantage of using directional flooding over general flooding is that packets being forwarded only in a single direction produces less overhead. The following proposition gives an estimate of the overhead due to directional flooding in terms of number of control packets.

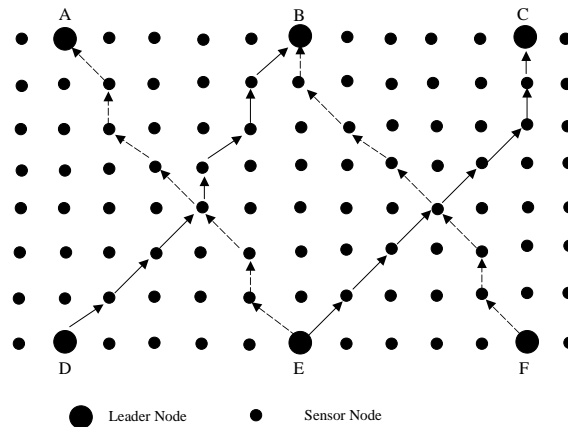


Figure 5.4: Simulation Topology

**Proposition 11** *In a geographically routed wireless sensor network with arbitrary topology and  $V$  nodes participating in the reverse flooding phase, exactly  $V$  control packets are transmitted.*

**Proof:** Note that  $V$ , the number of nodes participating in reverse flooding is expected to be  $\ll N$ , the total number of sensor nodes in the network. During the reverse flooding phase, at each node, only one packet becomes the winner among all the packets received by the node from its neighbors. All other packets are discarded. Each node then broadcasts the same packet to all of its neighbors that are in the geographic direction of the source node. Thus, a total of  $V$  nodes will send at most  $V$  broadcast packets. ■

Note that in a mesh topology, the actual number of control packets transmitted during the path determination phase can be reduced further. Consider the following situation shown in Figure 5.3. In this portion of the mesh, node A sends control packets to nodes B, C and D. If B receives the packet from A before its timer expires and if the energy value on this packet is the winning value at B as well, then B does not need to send this packet to node C (which has already received a copy of the same packet from A). In this manner, the total number of control packets can be further reduced.

## 5.6 Performance Evaluation

The main objective of our protocols is to gradually balance unfair energy consumption across the network and improve network lifetime as a consequence. However, there is no precise metric which measure effectiveness of energy balancing over network lifetime. To evaluate

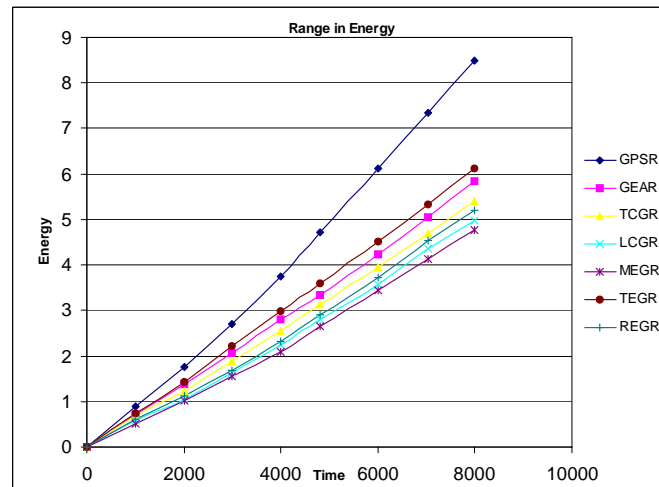


Figure 5.5: Range of residual energy levels across the network with GPSR, GEAR, LCGR, MEGR, REGR and TCGR

performance of our protocols, we use the following metrics which reflect dispersion of energy consumption across a network and an estimate of network lifetime in terms of percentage of completely exhausted nodes.

- Variance of energy level:** The variance of the energy levels of all the nodes is the primary measure of dispersion. A high variance indicates higher energy consumption at some of the nodes compared to others.
- Range of energy level:** This metric measures the difference between the energy levels of the maximum energy node and the minimum energy node over the whole network. A large value for this range is a result of unfair distribution of routing load among the nodes.
- Minimum energy node in the network:** This metric is related to the above two metrics. The energy value of the minimum energy node affects the variance and range of the energy values. A lower value of energy for the the minimum energy node indicates that the energy consumption is not balanced in the network.
- Time elapsed before a specified percentage of node dies:** This metric gives an estimate of network lifetime under the assumption that with more than 25 percent of completely exhausted nodes a network is no longer usable.

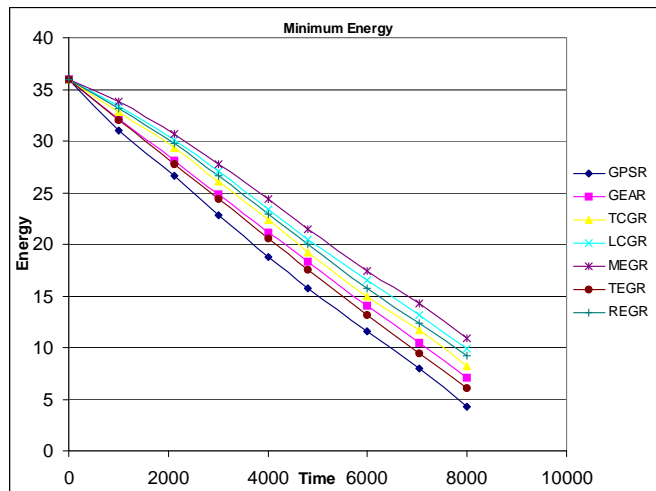


Figure 5.6: Minimum residual energy levels across the network with GPSR, GEAR, LCGR, MEGR, REGR and TCGR

### 5.6.1 Experimental Setup

In our simulation we have 104 nodes in a  $1000 \times 1000$  square meter area, with one node at each grid point of an  $8 \times 13$  square grid. Figure 5.4 represents the network topology we have used for evaluating our routing protocols. There are four pairs of source and destination nodes, namely (D,B), (E,A), (E,C) and (F,C). The underlying sensornet MAC protocol is TDMA based with source nodes D,E, and F generating packets at a uniform rate. Nodes D and F generate one packet every two frames. Node E generates a packet every frame with alternate packets destined to A and C.

We run the simulation for 8000 seconds and compare our five protocols LCGR, MEGR, TCGR, REGR and TEGR with each other and also with two well-known geographic routing protocols GPSR [83] and GEAR [86]. In GEAR, a node  $N_i$  dynamically chooses its minimum cost neighbor for forwarding its packet, where costs are parametrically estimated based on neighbors' geographical positions and energy levels as  $c(N_i) = \kappa d(N_i, R) + (1 - \kappa) E(N_i, R)$ . Here  $R$  is the target region and  $d(N_i, R)$  and  $E(N_i, R)$  are the neighborhood-normalized distance to destination and normalized energy consumed at  $N_i$  respectively [86]. In our simulation we use this routing strategy for source-to-destination communication with parameter  $\kappa$  set to 0.5 and compare it with our routing protocols.

These experiments are carried out on our simulation test-bed, which is an extension of Sensorsim [85]. The routing layer has been modified to implement our proposed protocols. In our simulations, reverse directional flooding is initiated in TCGR, LCGR and MEGR

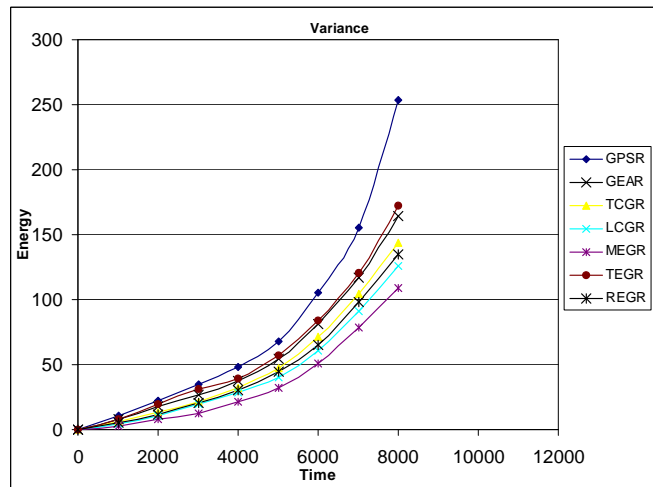


Figure 5.7: Variance in residual energy levels across the network under GPSR, GEAR, LCGR, MEGR, REGR and TCGR

when a destination node receives a sensor data packet indicating that at least *three* sensor nodes have energy values less than the threshold  $th$ . For the four protocols LCGR, MEGR, REGR and TEGR we use  $\beta = 0.9$ . For TCGR  $\alpha$  is set to 0.8 and  $\rho = 1.1$ .

### 5.6.2 Results and Analysis

We assume that before the network starts any activity, all ordinary sensor nodes have the same energy level. Therefore, in the very beginning, energy distribution is uniform across the network. When a network becomes active, the energy distribution across it gradually becomes non-uniform since nodes participating in a route inevitably consume more energy than other nodes. A protocol which uses a fixed path until one node in the route is completely drained results in a network energy distribution with high deviation. On the other hand, LCGR, MEGR, TCGR and REGR try to adapt to dynamically changing routes and gradually balance energy distribution across the network. Therefore, it is expected that the dispersion measures produced by our protocols will increase at a very slow rate with time. As a consequence, LCGR, MEGR, TCGR and REGR should result in longer network lifetime, where lifetime is defined in terms of a given percentage of nodes dying out.

Results of our simulation comparing performance of our protocols with that of GPSR and GEAR reflect this outcome. Figures 5.5, 5.6 and 5.7 show that all of our protocols produce smaller dispersion in energy levels. Figure 5.5 illustrates the range of node energy distributions across the network over time, under all four protocols. It can be seen that as

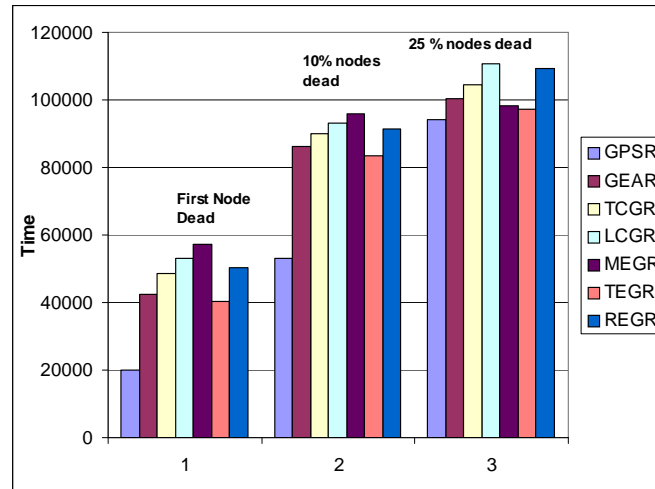


Figure 5.8: Time for different number of nodes to die in the network under GPSR, GEAR, LCGR, MEGR, TCGR and REGR

time proceeds, range of energy distribution produced under GPSR and GEAR increase at much higher rates compared to LCGR, MEGR and TCGR. Figure 5.6 shows that minimum residual energy level is higher under LCGR, MEGR, TCGR and REGR than under GEAR and GPSR. Similar results are achieved in Figure 5.7 when the variance of residual energy distribution is used as a performance metric. The higher variance under GPSR and GEAR indicates that a significant number of sensor nodes are being treated unfairly with network traffic being concentrated at fewer nodes. This might expedite partition of the network due to energy depletion at critical nodes. Under all the three metrics mentioned above, GPSR produces the worst performance since it uses a fixed route until a node's energy is completely drained out. GEAR performs better than GPSR because it dynamically changes routes using local network state information. Our protocols TCGR, LCGR, MEGR and REGR yield better performance than GEAR because they dynamically change routes using limited global knowledge combined with local information, whereas GEAR is a predominantly local protocol. Among the five proposed protocols, *for the metrics of energy range, minimum energy and deviation*, MEGR outperforms the others because it always uses the 'strongest' energy path (the path with least weakness) unconstrained by path length. Both LCGR and MEGR perform better than TCGR since TCGR produces less-optimal paths computed using threshold energy levels.

Figure 5.8 shows the time for specified percentages of nodes to die under all five protocols. According to this graph, our protocols perform better than both GEAR and GPSR.

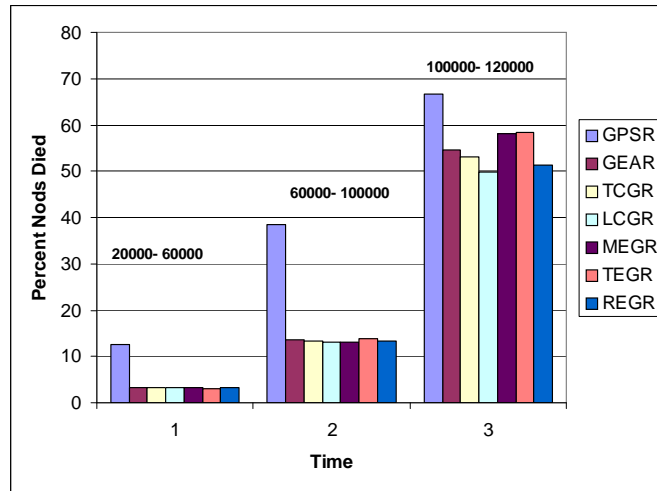


Figure 5.9: Percentage of nodes in the network that are dead at the end of specified intervals under GPSR, GEAR, LCGR, MEGR, TCGR and REGR

Although initially MEGR has the best performance, after a long period of time the number of nodes that die under MEGR is higher than TCGR, LCGR and REGR. This is because MEGR uses longer paths involving more number of nodes and producing higher overall energy consumption. Figure 5.9 shows the percentage of nodes dead at the end of different time intervals under the five protocols. This also indicates that performances of LCGR, TCGR and REGR become better with time compared to MEGR.

## 5.7 Conclusion

We describe a decision-theoretic paradigm for source-to-destination routing in a clustered sensor network architecture in which cluster heads utilize underlying network infrastructure for communication. The four protocols i.e., LCGR, MEGR, REGR and TCGR find length-energy-constrained optimal paths corresponding to the equilibrium of the route formation process. They also balance energy consumption across the network by selecting new optimal paths periodically. The simulation results indicate effectiveness of these protocols for enhancing network survivability.

# Chapter 6

## Summary of Thesis and Future Work

### 6.1 Summary of Thesis

The thesis titled as "Active Security Mechanisms for Wireless Sensor Networks and Energy Optimization for Passive Security Routing" focuses on various security aspects of wireless sensor networks deployed in a hostile environment and subject to enemy attack. We consider two types of security mechanism - active and passive. Active security mechanisms provide secure protocols, encryption methods and security hardware while passive security mechanisms in a sensor network aims at making conditions suitable for implementation of suitable active security mechanisms. We consider three different problems - the bootstrapping problem, secure data aggregation and energy-optimized routing. We review corresponding literature and provide solutions that outperform the existing solutions.

Establishing a secure communication infrastructure among a collection of randomly deployed sensor nodes is known as bootstrapping problem. Key pre-distribution is the most feasible and efficient solution for this problem. In this dissertation, we propose a novel solution to the key predistribution problem (labeled 2-Phase key predistribution) that exploits the connectivity and capture-resiliency properties of loading sensor nodes with a combination of *randomly derived* and *inherited* keys. We evaluate our solution by analytically developing novel quantitative metrics that measure the key predistribution schemes' security-performance tradeoffs in terms of the network resiliency to node/key capture, the number of available secure links and the key (memory) requirement per node for a given level of connectivity. We compare the network connectivity and security performance and show analytically and through simulations that the proposed 2-Phase scheme strongly favors highly secure large-composite key communication and is more resilient to node capture than the random scheme. We first show analytically that the invulnerability of an arbitrary  $q$ -composite

communication link to any number of node captures is higher in our scheme. We also derive analytical results for measuring the vulnerability of a  $q$ -composite link to single-node capture assuming adversaries who can use captured-key knowledge network-wide as well as locally and show that the 2-Phase scheme is more resilient. Finally, we present simulation results that show the number of exclusive keys shared between two nodes is higher while the number of  $q$ -composite links compromised when a given number of nodes are captured by the enemy is smaller under the 2-Phase scheme.

Secure data aggregation is another challenging research issue in wireless sensor networking. In this dissertation, we address the problem of enabling correct information aggregation, given that a fraction of the aggregators and ordinary sensors might be faulty or compromised. We divide the problem into two parts: a) a fraction of sensors are faulty/corrupt and an aggregator needs to detect inaccurate information; b) a fraction of aggregators are faulty/corrupt and the user needs to detect the corruption and accept only the true result with high probability. For the first part of the problem, we propose a novel probabilistic data accuracy model which enables an aggregator to estimate accuracy of each sensor reading by exploiting the fact that sensor readings are spatially correlated. We propose a novel algorithm which computes a weighted aggregate of data by attaching less weights to the sensor readings which are more likely to be wrong according to the proposed data accuracy model. We show by simulation that the weighted aggregate is more accurate than a simple aggregate. In the second part, we consider the problem of data aggregation when an aggregator is faulty/corrupted. We propose a multi-aggregator setup where a set of aggregators collect and aggregate data from the same region and send the information to the user. This enables the user to detect any corruption, thereby saving huge amount of energy. We propose several mechanisms which enable a user to accept the true aggregated value from an aggregator with correct readings with high probability.

Lastly, we consider the problem of energy-optimized routing which must distribute energy consumption uniformly across the network in order to prevent premature network partition. We model this problem of finding energy efficient routing paths with bounded lengths using a decision-theoretic paradigm in which routes are chosen using *profit* functions calculated at each sensor. The profit function is defined in such a way that it induces each node to link to the healthiest possible node while forming short paths. These two factors ensure that only a small number of relatively healthy nodes participate in routing, thereby reducing overall energy consumption and potentially delaying network partition. We define *length-energy-constrained* (LEC) optimal path as a path in which each sensor selects its best possible strategy given the choices of all other sensors. While computing this path is NP-hard in arbitrary sensor networks, we show that it can be found in polynomial time (in a distributed manner) in sensor networks operating under a geographic routing regime. However, the distributed algorithm requires a lot of state information to be stored in each sensor. We

propose a nearly stateless, scalable and easily implementable distributed protocol which approximately calculates an LEC optimal path. We also propose other protocols as well to calculate optimal paths under some variations of the proposed profit function. The first protocol *length-energy constrained geographic routing* (LCGR) uses a distributed algorithm to determine optimal paths defined by the above routing formation procedure in a geographically routed sensor network. The second protocol *max-min energy-constrained geographic routing* (MEGR) finds optimal paths for a simplified "team version" of the route formation process while the third protocol *threshold-energy constrained geographic routing* (TCGR) uses a threshold constrained heuristic for finding optimal paths.

## 6.2 Future Work

There is a broad variety of security aspects of wireless sensor networks that are still uncovered. While my long term goal is to explore those aspects for future research, there are immediate scopes of improving the solutions provided in this thesis.

- Two-phase key distribution:
  - While we have compared this scheme with the random key pre-distribution, we still do not know whether it outperforms the other schemes in the literature (e.g., pairwise key distribution schemes).
  - We have evaluated resiliency of our scheme to multiple node-capture only by simulation. In future, we aim at providing some analytical insight for this problem.
- Secure weighted data aggregation and accuracy verification:
  - We need to provide an analytical framework to evaluate the performance of the weighted data aggregation algorithm.
  - We did not provide any mechanism to determine the parameters involved in the accuracy model. We have evaluated them only by simulation. Although these parameters are application-specific, we need to provide some general guidelines so that the values of these parameters can be determined.
  - Solutions for detecting a compromised aggregator in a multi-aggregator setup depend on some strict assumptions on sensor reading. In future, we need to provide a more general solution without those assumptions.
  - We aim at providing a cost analysis of the algorithms presented in this thesis.

- Energy-optimized routing:
  - We need to address the complexity issues of the algorithms provided for length-energy-constrained routing.
  - In future, we want to provide analytical framework of energy savings by the algorithms presented here.

# Bibliography

- [1] N. Xu, S. Rangwala, K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, D. Estrin, "A Wireless Sensor Network for Structural Monitoring," *Proceedings of the ACM Conference on Embedded Networked Sensor Systems(Sensys04)* , November 2004.
- [2] J. Elson, D. Estrin, "Wireless Sensor Networks: A bridge to the Physical World", *Wireless Sensor Networks* , Kluwer, 2004.
- [3] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless Sensor Networks: A Survey, *Computer Networks (Elsevier) Journal*, pp. 393- 422, March 2002.
- [4] S. Kumar and D. Shepherd, SensIT: Sensor information technology for the warfighter, in *Proc. 4th Int. Conf. on Information Fusion*, 2001, pp. TuC1-3TuC1-9.
- [5] J. Corella, Tactical automated security system (TASS): Air force expeditionary security, *SPIE Conf. Unattended Ground Sensor Technologies and Applications*, Orlando, FL, 2003.
- [6] <http://www.llnl.gov/str/JulAug01/Hills.html>
- [7] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, J. Anderson. "Wireless sensor networks for habitat monitoring", *ACM International Workshop on Wireless Sensor Networks and Applications (WSNA '02)*, Atlanta, GA, September 2002.
- [8] H. Wang, J. Elson, L. Girod, D. Estrin, and K. Yao, "Target classification and localization in habitat monitoring", *In Proceedings of the IEEE ICASSP 2003*, Hong Kong, April 2003.
- [9] E. Biagioni, K. Bridges, "The application of remote sensor technology to assist the recovery of rare and endangered species", *Special issue on Distributed Sensor Networks for the International Journal of High Performance Computing Applications*, Vol. 16, N. 3, August 2002.

- [10] <http://www.alertsystems.org>.
- [11] L. Schwiebert, S. Gupta, J. Weinmann, "Research challenges in wireless networks of biomedical sensors", *In Mobile Computing and Networking*, pages 151. 165, 2001.
- [12] E. Shih et al., "Physical Layer Driven Protocol and Algorithm Design for Energy Efficient Wireless Sensor Networks," *Proc. ACM Mobicom '01*, Rome, Italy, July 2001, pp. 272-286.
- [13] A. Woo, D. Culler, "A Transmission Control Scheme for Media Access in Sensor Networks", *Proc. ACM Mobicom '01*, Rome, Italy, July 2001, pp. 221-235.
- [14] J.M. Kahn, R.H. Katz, K.S.J. Pister, "Next Century Challenges: Mobile Networking for Smart Dust", *Proc. ACM Mobicom '99*, Washington D.C., 1999, pp.271-78.
- [15] G.J.Pottie, W.J. Kaiser, "Wireless Integrated Network Sensors", *Comm. ACM*, Vol.43, No.5, May 2000, pp.551-58.
- [16] Soo-young Shin, "Performance Evaluation of Interference Between Bluetooth Networks using Bit Error Rate", *Proceedings on the 15th CISL Winter Workshop*, Kushu, Japan, February 2002.
- [17] P. Papadimitratos, Z. J. Haas, "Secure Routing for Mobile Ad hoc Networks", *Proceedings of the SCS Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS 2002)*, San Antonio, TX, January 2002.
- [18] M.J. Handy, M. Haase and D. Timmermann, "Low Energy Adaptive Clustering Hierarchy with Deterministic Cluster-Head Selection," *4th IEEE International Conference on Mobile and Wireless Communications Networks*, Stockholm, 2002.
- [19] C. Intanagonwiwat, Ramesh Govindan and Deborah Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," *In Proceedings of the Sixth Annual International Conference on Mobile Computing and Networks (MobiCOM 2000)*, August 2000, Boston, Massachusetts.
- [20] P. Papadimitratos, Z. J. Haas, "Secure Routing for Mobile Ad hoc Networks", *Proceedings of the SCS Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS 2002)*, San Antonio, TX, January 2002.
- [21] Z. Huang, C. Shen, "A comparison study of omnidirectional and directional MAC protocols for ad hoc networks", *IEEE Global Telecommunications Conference*, Vol. 1, Nov. 2002, pp. 57 - 61.

- [22] Alaa Muqattash and Marwan Krunz, "CDMA-based MAC protocol for wireless ad hoc networks", *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking and computing*, June 2003, pp. 153 - 164.
- [23] Tiantong You and Hossam Hassanein, "Infrastructure-based MAC in wireless mobile ad-hoc networks", *27th Annual IEEE Conference on Local Computer Networks*, pp. 821 - 830, November 2002.
- [24] Tianbo Kuang, Carey Williamson, "A bidirectional multi-channel MAC protocol for improving TCP performance on multihop wireless ad hoc networks", *Proceedings of the 7th ACM international symposium on Modeling , analysis and simulation of wireless and mobile systems*, pp. 301 - 310, 2004.
- [25] Chunhung Richard Lin and Mario Gerla, "Adaptive clustering for mobile wireless networks", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 7, pp. 1265 - 1275, September 1997.
- [26] S. Singh and C.S. Raghavendra, PAMAS: Power aware multi-access protocol with signalling for ad hoc networks, *ACM Computer Communication Review*, vol. 28, no. 3, pp. 526, July 1998. W. Ye,
- [27] J. Heidemann, and D. Estrin, An energy-efficient mac protocol for wireless sensor networks, *Proc. IEEE INFOCOM, New York, NY*, June 2002, pp. 15671576.
- [28] Katayoun Sohrabi and Gregory J. Pottie, Performance of a novel selforganization protocol for wireless ad hoc sensor networks, *Proceedings of the IEEE 50th Vehicular Technology Conference*, 1999, pp. 12221226.
- [29] Frazer Bennett, David Clarke, Joseph B. Evans, Andy Hopper, Alan Jones, and David Leask, Piconet: Embedded mobile networking, *IEEE Personal Communications Magazine*, vol. 4, no. 5, pp. 815, Oct. 1997.
- [30] Alec Woo and David Culler, A transmission control scheme for media access in sensor networks, *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking*, Rome, Italy, July 2001, ACM.
- [31] Adam Dunkels, Juan Alonso, Thiemo Voigt, "Making TCP/IP Viable for Wireless Sensor Networks", <http://www.sics.se/adam/ewsn2004.pdf>
- [32] W.R. Heinzelman, J. Kulik, H. Balakrishnan, "Adaptive Protocols for Information Dissemination in Wireless Sensor Networks", *Proc. ACM MobiCom '99*, Seattle, WA, 1999, pp. 174-85.

- [33] S. Hedetiniemi, A. Liestman, "A Survey of Gossiping and Broadcasting in Communication Networks", *Networks*, vol. 18, 1988.
- [34] K. Sohrabi et al., "Protocols for Self-Organization of a Wireless Sensor Network", *IEEE Pers. Communication*, Oct 2000, pp. 16-27.
- [35] G.J. Pottie, W.J. Kaiser, Wireless Integrated Network Sensors, *Comm. ACM*, vol. 43, no. 5, May 2000, pp. 551-58.
- [36] K. Sohrabi et al., Near-ground Wideband Channel Measurements, *IEEE Proc. VTC*, New York, 1999.
- [37] C. Chien et al., Low-power Direct-sequence Spread Spectrum Modern Architecture for Distributed Wireless Sensor Networks, *ISLPED '01*, CA, Aug 2001.
- [38] H.Chan, A.Perrig and D.Song, "Random Key Pre-distribution Schemes For Sensor Networks", *IEEE Symposium on Security and Privacy*, 2003.
- [39] D. Estrin, L. Girod, G. Pottie and M. Srivastava, "Instrumenting the world with wireless sensor networks," *Proceedings of the 10th Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, July 2001.
- [40] D. Estrin and R. Govindan, "Next Century Challenges: Scalable Coordination in Sensor Networks," in *Proc. ACM/IEEE Conf. Mobicom'99*, pp. 263-270, Aug. 1999.
- [41] L.Eschenaur and V.D.Gligor, "A Key-management Scheme For Distributed Sensor Networks", *Proceedings of the 9th ACM Conference on Computer and Communication Security*, pp.41-47, November 2002.
- [42] R. W. Heinzelman, J. Kulik and H. Balakrishnan, "Adaptive Protocols for Information Dissemination in Wireless Sensor Networks," in *Proc. ACM/IEEE Conf. Mobicom'99*, pp. 174-185, Aug. 1999.
- [43] S. Zhu, S. Setia, and S. Jajodia, "LEAP: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks. *Proc. of the 10th ACM Conference on Computer and Communications Security (CCS '03)*, Washington D.C., October, 2003
- [44] S. Zhu, S.Xu, S. Setia and S. Jajodia, "Establishing Pair-wise Keys For Secure Communication in Ad Hoc Networks: A Probabilistic Approach," *11th IEEE International Conference on Network Protocols (ICNP'03)*, Atlanta, Georgia, November 4-7, 2003.

- [45] R. Kalidindi, V.Parachuri, R.Kannan, A. Durresi and S. Iyengar, "Sub-Quorum Based Key Vector Assignment: A Key Pre-Distribution Scheme For Wireless Sensor Networks," *Intl. Conf. On Wireless Networking, (ICWN04)*, Las Vegas, July 2004.
- [46] C. Karlof and D. Wagner, "Secure Routing in Sensor Networks: Attacks and Countermeasures", *First IEEE International Workshop on Sensor Network Protocols and Applications*, May 11, 2003 Anchorage, AK, USA
- [47] D. Liu and P. Ning, "Establishing Pairwise Keys in Distributed Sensor Networks," *10th ACM Conference on Computer and Communications Security (CCS '03)*, Washington D.C., October, 2003.
- [48] M. Naor and A. Wool, "The Load Capacity and Availability of Quorum Systems," *SIAM J. on Computing*, April 1998.
- [49] A.Perrig, R.Szewczyk, V.Wen, D.Culler and J.D. Tygar, "Spins: Security Protocols for Sensor Networks", *Proc. of 7th Int'l Distributed Sensor Networks*, pp. 163–168, Aug. 1998.
- [50] S. Slijepcevic, V. Tsiatsis, S. Zimbeck, M. B. Srivastava, and M.Potkonjak, "On Communication Security in Wireless Ad-Hoc Sensor Networks," *11th IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET-ICE)*, Pittsburgh, June 20 02, pp. 139-144.
- [51] "SMART DUST: Autonomous sensing and communication in a cubic millimeter," <http://robotics.eecs.berkeley.edu/pister/SmartDust>
- [52] J.Spencer, "The Strange Logic of Random Graphs", *Algorithms and Combinatorics 22*, Springer-Verlag 2000, ISBN 3-540-41654-4.
- [53] K. Sohrabi, J. Gao, V. Ailawadhi and G. Pottie, "Protocols for Self-Organization of a Wireless Sensor Network," *IEEE Personal Comm. Magazine*, Vol. 7, No. 5, pp. 16-27, October 2000.
- [54] A. Wood and J. Stankovic, "Denial of Service in Sensor Networks," *IEEE Computer*, pp. 54–62, October 2002.
- [55] L. Zhou and Z. Haas, "Securing Ad-Hoc Networks," *IEEE Network Magazine*, Vol. 13, No. 6, pp. 24-30, 1999.

- [56] J. Zhao and R. Govindan, "Understanding packet delivery performance in dense wireless sensor networks", *In Proceedings of the International Conference on Embedded Networked Sensor Systems (SenSys 2003)*, November 2003.
- [57] J. Kohl and B. Clifford Neuman, "The Kerberos Network Authentication Service (V5)", *RFC 1510*, September 1993.
- [58] S. P. Miller, C. Neuman, J. I. Schiller, and J. H. Saltzer, "Kerberos authentication and authorization system", *Project Athena Technical Plan*, page section E.2.1, 1987.
- [59] David W. Carman, Peter S. Kruus, and Brian J. Matt, "Constraints and approaches for distributed sensor network security", *NAI Labs Technical Report no.00-010*, September 2000.
- [60] W. Diffie, M.E. Hellman, "New Directions in Cryptography", *IEEE Transactions on Information Theory*, vol.22, November 1976, pp. 644-654.
- [61] R.L. Rivest, A. Shamir, L.M. Adleman, "A method for Obtaining Digital Signatures and Public Key Cryptosystems", *Communications of ACM*, vol.21, no.2, 1978, pp. 120-126.
- [62] R. Merkle, "Secure communication over insecure channels", *Communications of the ACM*, vol.21, no. 4, pp. 294-299, 1978.
- [63] R. Blom, "Non-public key distribution", *Advances in Cryptology: Proceedings of Crypto 82*, pp. 231-236, 1982.
- [64] C. Blundo, A. De Santis, A. Herzberg, S. Kutten, U. Vaccaro, and M. Yung, "Perfectly-secure key distribution for dynamic conferences" *Advances in Cryptology - Crypto 92*, pp. 471-486, 1992
- [65] Wenliang Du, Jing Deng, Yunghsiang S. Han, and Pramod K. Varshney, "A pairwise key pre-distribution scheme for wireless sensor networks", *ACM CCS 2003*, pp. 425-431, October 2003.
- [66] Donggang Liu and Peng Ning, "Establishing pairwise keys in distributed sensor networks", *ACM CCS 2003*, pp. 526-531, October 2003.
- [67] B.Krishnamachari, D.Estrin, S.Wicker, "The Impact of Data Aggregation in Wireless Sensor Networks", *In International Workshop of Distributed Event Based Systems (DEBS)*, Austria, July 2002.

- [68] C.Intanagonwiwat, D.Estrin, R.Govindan, J.Heidemann, "Impact of network density on data aggregation in wireless sensor networks", *In Proceedings of International Conference on Distributed Computing Systems (ICDCS)*, Vienna, Austria, July 2002.
- [69] B.Przydatek, D.Song, A.Perrig, "SIA: secure information aggregation in sensor networks", *Proceedings of the first international conference on Embedded networked sensor systems* , pp.255-265, 2003.
- [70] A.Perrig, J.Stankovic, D.Wagner, "Security in wireless sensor networks ", *Communications of the ACM*, 47(6), June 2004, pp.53-57.
- [71] B. Krishnamachari and S. Iyengar, "Bayesian Algorithms for Fault-tolerant Event Region Detection in Wireless Sensor Networks," *IEEE Transactions on Computers*, 2004.
- [72] C.Karlof, D.Wagner, "Secure Routing in Wireless Sensor Networks: Attacks and Countermeasures", *First IEEE International Workshop on Sensor Network Protocols and Applications*, May 11, 2003.
- [73] A.M.Goon, M.K.Gupta, B.Dasgupta, *An Outline of Statistical Theory*, Vol.2, The World Press Private Limited, India.
- [74] H. Robbins, "Sequential Estimation of a Normal Mean Population", *Probability and Statistics - The Harald Cramer Volume*", Uppsala, 1959, pp. 233-245.
- [75] Samuel R. Madden, Michael J. Franklin, Joseph M. Hellerstein, and Wei Hong, "TAG: a Tiny AGgregation service for ad-hoc sensor networks", *Proceedings of the Fifth Annual Symposium on Operating Systems Design and Implementation (OSDI)*, December 2002.
- [76] Amol Deshpande, Suman Nath, Phillip B. Gibbons, and Srinivasan Seshan, "Cache-and-query for wide area sensor databases", *SIGMOD 2003*, 2003.
- [77] Lingxuan Hu and David Evans, "Secure aggregation for wireless networks", *Workshop on Security and Assurance in Ad hoc Networks*, January 2003.
- [78] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem", *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol.4, no.3, pp.382-401, July 1982.
- [79] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, 1991.
- [80] D.B. Johnson and D.A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Sensor Networks", *Mobile Computing*, Kluwer, 1996.

- [81] C. Perkins and E. Royer, "Ad Hoc On Demand Distance Vector Routing", *Proc. of 2nd IEEE Wksp Mobile Comp. Sys. and Applications*, February 1999.
- [82] J. Chang and L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks," *IEEE INFOCOM '2000*, March 2000.
- [83] Brad Karp and H. T. Kung, "GPSR: Greedy perimeter stateless routing for wireless networks," *Proc. ACM/IEEE MobiCom*, August 2000.
- [84] Rahul C. Shah and Jan M. Rabaey, "Energy Aware Routing For Low Energy Ad Hoc Sensor Networks," *Proc. IEEE WCNC'02*, March 2002.
- [85] S. Park, A. Savvides and M. B. Srivastava, "SensorSim: A Simulation Framework for Sensor Networks," *Proceedings of MSWiM 2000*, Boston, MA, August 11, 2000.
- [86] Yan Yu, Ramesh Govindan and Deborah Estrin, "Geographical and Energy Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks," *UCLA Computer Science Department Technical Report UCLA/CSD-TR-01-0023*, May 2001.
- [87] K. Seada, M. Zuniga, A. Helmy, B. Krishnamachari, "Energy-Efficient Forwarding Strategies for Geographic Routing in Lossy Wireless Sensor Networks", *SenSys 2004*, November, 2004, Baltimore, Maryland.
- [88] R. Kannan, L. Ray, R. Kalidindi, and S.S. Iyengar, "Threshold-Energy Constrained Protocol for Wireless Sensor Networks," *Sensor Processing Letters*, vol 1, December 2003.
- [89] S. Lindsey, C. Raghavendra and K. Sivalingam, "Data Gathering in Sensor Networks using the Energy Delay Metric", *In International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing*, Apr. 2001, San Francisco, CA.
- [90] J. Kulik, W. R. Heinzelman, and H. Balakrishnan, "Negotiation-based Protocols for Disseminating Information in Wireless Sensor Networks", *ACM/IEEE Int. Conf. on Mobile Computing and Networking*, Aug. 1999, Seattle, WA.
- [91] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", *In Proceedings of the Hawaii Conference on System Sciences*, January 2000.

# Vita

The author of this thesis, Lydia Ray, is a native of West Bengal, India. Born on 19th December in 1975, she received a bachelor's degree (B.Sc) with honours in statistics from Presidency College under Calcutta University. She received a master's degree (M.Stat) in Statistics from Indian Statistical Institute in 1998 with a first class. After pursuing research as a research assistant in Indian Statistical Institute for two years, she decided to change the field of study. She came to University of Alabama in Huntsville in Fall 2000 as a doctoral student. She joined Louisiana State University as a doctoral student in the Department of Computer Science in Fall 2001. She conducted research on wireless sensor networks under the guidance of Dr. Rajgopal Kannan and completed her dissertation in Spring 2005. She took her final examination in April, 2005 and will receive the degree of Doctor of Philosophy in August, 2005.