DISTRIBUTED SPATIAL DATA AGGREGATION AND DILUTION BASED ON HASHING AND RELATIONAL ALGEBRA IN WIRELESS SENSOR NETWORKS

M.S. Thesis by

Tolga ÇÖPLÜ

704011007

Date of submission : 26 April 2004
Date of defence examination : 07 June 2004

Supervisor (Chairman): Assoc. Prof. Sema OKTUĞ
Members of the Examining Committee: Assoc. Prof. Erdal ÇAYIRCI (Co-Supervisor)
   Prof. Dr. Emre Harmancı
   Prof. Dr. Bülent Örencik
   Prof. Dr. Şebnem Baydere

MAY 2004
PREFACE

I would like to extend my gratitude to my supervisors Assoc. Prof. Sema OKTUĞ and Assoc. Prof. Erdal ÇAYIRCI for their valuable guidance and contributions. I would also like to express my appreciation to Assoc. Prof. Erdal ÇAYIRCI for sharing his studies.

May, 2004                      Tolga Çöplü
## CONTENTS

PREFACE .................................................................................................................. ii
CONTENTS ................................................................................................................ iii
ABBREVIATIONS .................................................................................................... iv
LIST OF FIGURES ................................................................................................... v
ABSTRACT ................................................................................................................ vi
ÖZET ........................................................................................................................ vii

1 INTRODUCTION ................................................................................................. 1
  1.1 Contribution of the Thesis ............................................................................... 5
  1.2 Structure of the Thesis .................................................................................... 5

2 RELATED WORK .................................................................................................. 7
  2.1 Data Querying in Sensor Networks .................................................................. 7
  2.2 Data Aggregation in Sensor Networks ............................................................ 8
  2.3 Routing in Sensor Networks .......................................................................... 9

3 DATA AGGREGATION and DILUTION in SENSOR NETWORKS .............. 12
  3.1 System Architecture ....................................................................................... 12
  3.2 Querying Sensor Network in DADMA ............................................................ 14
  3.3 Implementation of DADMA .......................................................................... 17

4 PERFORMANCE EVALUATION ......................................................................... 23
  4.1 The Percentage of Diluted Nodes ................................................................. 24
  4.2 Probability of Detection ............................................................................... 25
  4.3 The Probability of Aggregation .................................................................... 30

5 EXPERIMENTAL RESULTS ............................................................................. 32

6 CONCLUSION AND FUTURE WORK .............................................................. 50
REFERENCES ......................................................................................................... 52
BIOGRAPHY ............................................................................................................ 55
# Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSN</td>
<td>Wireless Sensor Networks</td>
</tr>
<tr>
<td>DADMA</td>
<td>Data Aggregation and Dilution by Modulus Addressing</td>
</tr>
<tr>
<td>VLSNT</td>
<td>Virtual Local Sensor Node Table</td>
</tr>
<tr>
<td>SNDV</td>
<td>Sensor Network Database View</td>
</tr>
<tr>
<td>ESNDT</td>
<td>External Sensor Network Database Table</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SPIN</td>
<td>Sensor Protocols for Information via Negotiation</td>
</tr>
<tr>
<td>LEACH</td>
<td>Low-Energy Adaptive Clustering Hierarchy</td>
</tr>
<tr>
<td>SMACS</td>
<td>Self-Organizing Medium Access Control for Sensor Networks</td>
</tr>
<tr>
<td>ACQUIRE</td>
<td>Active Query Forwarding</td>
</tr>
<tr>
<td>MARC</td>
<td>Mobility Assisted Resolution of Queries</td>
</tr>
<tr>
<td>SAR</td>
<td>Sequential Assignment Routing</td>
</tr>
<tr>
<td>PEGASIS</td>
<td>Power Efficient Gathering in Sensor Information Systems</td>
</tr>
<tr>
<td>MECN</td>
<td>Minimum Energy Communication Network</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>SPIN protocol</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Directed Diffusion</td>
<td>11</td>
</tr>
<tr>
<td>3.1</td>
<td>A sensor network perceived as a distributed database</td>
<td>14</td>
</tr>
<tr>
<td>3.2</td>
<td>The structure of an SQL statement for DADMA</td>
<td>15</td>
</tr>
<tr>
<td>3.3</td>
<td>Dilute-m, resolution $r$</td>
<td>16</td>
</tr>
<tr>
<td>3.4</td>
<td>Aggregation in Sensor Field</td>
<td>17</td>
</tr>
<tr>
<td>3.5</td>
<td>The DADMA Algorithm</td>
<td>18</td>
</tr>
<tr>
<td>3.6</td>
<td>An illustrative example for the implementation of DADMA</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>Sensor field</td>
<td>23</td>
</tr>
<tr>
<td>4.2</td>
<td>Sensor field when aggregate $m$ function is used</td>
<td>30</td>
</tr>
<tr>
<td>5.1</td>
<td>Percentage of diluted nodes 1</td>
<td>33</td>
</tr>
<tr>
<td>5.2</td>
<td>Percentage of diluted nodes 2</td>
<td>33</td>
</tr>
<tr>
<td>5.3</td>
<td>Percentage of diluted nodes 3</td>
<td>34</td>
</tr>
<tr>
<td>5.4</td>
<td>Percentage of diluted nodes 4</td>
<td>35</td>
</tr>
<tr>
<td>5.5</td>
<td>Percentage of diluted nodes 5</td>
<td>35</td>
</tr>
<tr>
<td>5.6</td>
<td>Percentage of diluted nodes 6</td>
<td>36</td>
</tr>
<tr>
<td>5.7</td>
<td>Probability of detection 1</td>
<td>37</td>
</tr>
<tr>
<td>5.8</td>
<td>Probability of detection 2</td>
<td>37</td>
</tr>
<tr>
<td>5.9</td>
<td>Probability of detection 3</td>
<td>38</td>
</tr>
<tr>
<td>5.10</td>
<td>Probability of detection 4</td>
<td>38</td>
</tr>
<tr>
<td>5.11</td>
<td>Probability of aggregation 1</td>
<td>39</td>
</tr>
<tr>
<td>5.12</td>
<td>Probability of aggregation 2</td>
<td>40</td>
</tr>
<tr>
<td>5.13</td>
<td>Number of transmitted packets 1</td>
<td>41</td>
</tr>
<tr>
<td>5.14</td>
<td>Number of transmitted packets 2</td>
<td>41</td>
</tr>
<tr>
<td>5.15</td>
<td>Number of transmitted packets 3</td>
<td>42</td>
</tr>
<tr>
<td>5.16</td>
<td>Number of transmitted packets 4</td>
<td>42</td>
</tr>
<tr>
<td>5.17</td>
<td>Number of transmitted packets 5</td>
<td>43</td>
</tr>
<tr>
<td>5.18</td>
<td>Number of transmitted packets 6</td>
<td>43</td>
</tr>
<tr>
<td>5.19</td>
<td>Number of transmitted packets 7</td>
<td>44</td>
</tr>
<tr>
<td>5.20</td>
<td>Number of transmitted packets 8</td>
<td>44</td>
</tr>
<tr>
<td>5.21</td>
<td>Number of transmitted packets 9</td>
<td>45</td>
</tr>
<tr>
<td>5.22</td>
<td>Number of transmitted packets 10</td>
<td>46</td>
</tr>
<tr>
<td>5.23</td>
<td>Number of transmitted packets 11</td>
<td>46</td>
</tr>
<tr>
<td>5.24</td>
<td>Number of transmitted packets 12</td>
<td>47</td>
</tr>
<tr>
<td>5.25</td>
<td>Number of transmitted packets 13</td>
<td>47</td>
</tr>
<tr>
<td>5.26</td>
<td>Number of transmitted packets 14</td>
<td>48</td>
</tr>
<tr>
<td>5.27</td>
<td>Number of transmitted packets 15</td>
<td>48</td>
</tr>
<tr>
<td>5.28</td>
<td>Number of transmitted packets 16</td>
<td>49</td>
</tr>
</tbody>
</table>
DISTRIBUTED SPATIAL AGGREGATION AND DILUTION BASED ON HASHING AND RELATIONAL ALGEBRA IN WIRELESS SENSOR NETWORKS

ABSTRACT

In my thesis, a new scheme is introduced to effectively query sensor nodes in wireless ad-hoc sensor networks. Wireless sensor networks are based on collaborative effort of sheer number of tiny sensor nodes deployed either close to or inside the phenomenon to be observed. We perceive wireless sensor networks as a distributed database and based on this perception an effective data query scheme is employed where users interact with the network by using a standard SQL like statements. In this scheme, a new algorithm that can run on tiny sensor nodes to aggregate or dilute the sensed data packets is used. Two location based hash functions are also introduced to determine how the sensed data can be grouped or which sensors should be excluded from a query. Analytical models are provided for the performance evaluation. The numerical results show that the proposed scheme can reduce the number of transmitted packets 50% on the average comparing to the case where aggregation or dilution is not used.
TELSİZ SENSÖR AĞLARDA ÖZALMA VE İLİŞKİSEL CEBİR KULLANILARAK DAĞITILmiş YÖРЕSEL VERİ BÜTÜНLEŞTIРME VE SÜZMESİ

ÖZET

Bu tez çalışmasında telsiz ad-hoc sensör ağlardaki düğümlerin efektif bir şekilde sorgulanmasını sağlayan bir yöntem geliştirilmiştir. Telsiz sensör ağlar, işlem ve haberleşme kapasitesi düşük, gözlenecek olayın yakınına ya da içine konuşlandırılmış çok saydaki sensör düğümlerinin iş birliği yaparak çalışması prensibine dayanır. Sensör ağı çalıştırılmasında dağıtılmış bir veritabanı olarak soyutlanmıştır ve kullanıcıların SQL benzeri standart bir dille ağa erişip veri sorgulaması yapmasına olanak sağlanmıştır. Yapıda sensör düğümleri üzerinde çalışabilecek ve veri paketlerinin bütünleştirilmesi veya süzülmesini sağlayan yeni bir algoritma geliştirilmiştir. İki adet özalma fonksiyonu önerilerek, yöresel olarak ölçülen verinin bütünleştirilmesi veya bazı düğümlerin ölçüme işleminden dışlanması mümkün kılınmıştır. Yöntemin performans değerlendirilmesinin yapılabileceği için analitik modeller oluşturulmuş ve simülasyonlar yapılmıştır. Sonuçlar yöntemin bütünleştirme veya süzme kullanılmayan durumlara göre iletilen paket sayını ortalama %50 azalttığını göstermektedir.
1 INTRODUCTION

The developments in wireless networking technologies and the continuous decrease in the size and the cost of electro mechanical systems [1,27] have born the idea of rapidly deployable, self-organized network of tiny sensor nodes and actuators. This idea has created a great research interest in pervasive sensing and control. Based on this interest, we can say that wireless sensor and actuator networks will become an integrated part of our daily lives soon. The concept of tiny sensor nodes supply significant advantages compared to traditional sensors. Some of these advantages are [1,6,34]:

- Traditional sensors are expensive hardware’s and dependent to continuous energy supplies.
- Their deployment is a long time work and their locations are static.
- Since the number of sensors in a sensor network is much higher, a cooperative effort increases the throughput compared to traditional sensors.

Sensor networks seem to be a special type of ad-hoc networks due to their infrastructureless architecture. However, there are a serious number of factors that make sensor networks to be classified as a new communication architecture [1]. Some of these are:

- The amount of sensor nodes in a sensor network is much more than nodes in a ad-hoc network. Since scalability is one of the main problems of ad-hoc communication, the importance of this factor can be realized.
- Sensor nodes have more stringent energy computation and communication constraints compared to nodes in ad-hoc networks, because of their size and cost.
- Deployment [33,34] is mostly random in sensor networks and density of nodes is much higher.
• Nodes are prone to failure due to random deployment, physical factors and limited energy sources.

• Unique global identifications can not be used in many sensor network applications.

• Sensor network topologies are much more unstable.

• Mostly data centric routing is used in sensor networks.

To understand this new architecture clearly, let us examine the design characteristics of basic elements of sensor networks. Sensor nodes have 4 main parts [1], which are 1) processing unit, 2) communication unit, 3) sensing units, and 4) power unit. According to the type of application, some additional units can also be integrated to this structure, like a mobilizer, a power generator, or a location finding unit. Since we are dealing with great amount of sensor nodes for different deployment conditions, these hardwares must be small in size and low in price. Sensor nodes have limited processing, communication, energy and storage capabilities. Addition to these factors, effects of wireless medium make reliability, fault tolerance, scalability and energy efficiency the key problems of wireless sensor networks [1,35]. As a result new MAC, network, transport and application layer protocols that fit these characteristics must be developed [29].

In spite of the difficulties of this architecture, large number of application areas makes sensor networks an attractive concept. Also some recent research projects and applications, like MICA [26], TinyOS [27], Habitat Monitoring [28], Smart Dust [22], PicoRadio [23], SQTL[24], µAMPS [30], proved that sensor networks will soon be a part of our daily life.

Some of the sensor network application areas are:

• Military

• Health

• Chemical processing

• Environmental, disaster relief

• Industry

• Smart home applications
Since the lifetime of a wireless sensor network (WSN) is generally dependent on irreplaceable power sources in tiny sensor nodes, power efficiency is one of the critical design factors for WSNs [1,22,35]. The research of Pottie and Kaiser showed that the cost of transmitting 1 Kb of data to a distance of 100 meters is approximately equal to the cost of executing 3 million CPU instructions [18]. At this point, local processing and data aggregation [3,4,25,32] gains great importance as a particularly useful paradigm for routing in wireless sensor networks. By eliminating redundant packets and combining others coming from different paths, aggregation saves energy. Data aggregation techniques [16,17,31] that reduce the number of data packets conveyed through the network are therefore important and also required for effective fusion of data collected by vast number of sensor nodes [1][2]. Data aggregation in sensor networks unites the sensed data coming from the nodes based on the parameters passed in queries. It can be classified according to one of the following approaches:

**Temporal or spatial aggregation:** Data can be aggregated based on time or location. For example, the temperature readings taken every hour or temperature readings from various regions in a sensor field can be averaged. Also a hybrid approach which is the combination of time and location based aggregation can be used.

**Snapshot or periodical aggregation:** Data aggregation can be made snapshot, i.e., one time, on the receipt of a query. Alternatively, temporarily aggregated data can be reported periodically [16].

**Centralized or distributed aggregation:** A central node can gather and then aggregate data or data can be aggregated while being conveyed through a sensor network. A hybrid approach is also possible where clusters are formed, and a node in each cluster aggregates the data from the cluster.

**Early or late aggregation:** Data can be aggregated at the earliest opportunity, or aggregation of data may not be allowed before a certain number of hops not to hinder the collaboration among the neighboring nodes.

Data querying is an important phase of data aggregation because the rules that describe how to aggregate the sensed data are passed to the sensor network by queries. It should be noted that although a query is a part of a data aggregation
scheme, it can be made not only for aggregated data but also for non-aggregated data. A query in sensor network may be perceived as the task or interest dissemination process. Sensor nodes can be queried by using continuous or snapshot queries. Continuous queries can be periodical where the sensed data are reported at certain time intervals or event driven where certain events stimulate nodes to report the sensed data.

The following characteristics of WSNs should be considered while designing an aggregation or data querying scheme.

- Sensor nodes are limited in both memory and computational resources. They cannot buffer large number of data packets.

- Sensor nodes generally disseminate short data packets to report an ambient condition, e.g., temperature, pressure, humidity, proximity report, etc.

- The observation areas of sensor nodes often overlap. Therefore, many sensor nodes may report the correlated data related to the same event. However, in many cases the replicated data are needed because sensor network concept is based on the cooperative effort of low fidelity sensor nodes [1]. For example, nodes may report only proximity, and then the size and the speed of the detected object can be derived from the locations of the nodes reporting them, and timings of the reports. The collaboration among the nodes should not be hampered by the data aggregation scheme.

- Since there may be thousands of nodes in a sensor field, associating data packets from numerous sensors to the corresponding events, and correlating the data about the same event reported at different times may be very complicated task for a single sink node or a central system.

- Due to large number of nodes and other constraints such as power limitations, sensor nodes are not globally addressed [1]. Therefore, only address-centric protocols (end-to-end routing) is mostly inefficient. Instead of address-centric protocols, data-centric or location aware addressing protocols where intermediate nodes can route data according to its content [6] or the location of the nodes [12], should be used.

- Querying the whole network node by node is impractical. So attribute-based naming and data-centric routing [7] are essential for WSNs.
1.1 Contribution of the Thesis

In this thesis, a new distributed spatial data aggregation and dilution scheme based on hashing and relational algebra is introduced for wireless sensor networks to fit the above characteristics. This work is mainly focused on querying sensor networks and satisfying energy efficiency by aggregation and/or dilution. We introduce a practical distributed algorithm to aggregate the sensed data. This algorithm is simple enough to run on sensor nodes. It does not require the time synchronization of nodes, supports node mobility and allows controlling sensor off duty cycles by using a middleware architecture between application and network layers. Besides, data aggregation and dilution by modulus addressing (DADMA) can be used in conjunction with the known routing schemes such as Directed Diffusion [3], SPIN [2], SMACS [5] and LEACH [10].

A WSN is perceived as a distributed relational database that has a single view, which is created by joining records distributed in virtual local sensor node tables. Sensed data are retrieved from WSN by an SQL like statement. While data are being retrieved, they are also aggregated or diluted based on the rules defined in the query.

1.2 Structure of the Thesis

In Section 2, related work is studied in three subsections. First, proposed data querying methods for WSNs are summarized. Their deficiencies are described according to our perspective. Then data aggregation approaches in literature are surveyed and criticized. Finally, in the last subsection, common routing techniques for sensor networks are summarized.

In Section 3, the distributed relational database abstraction of DADMA is described in detail. After explaining the general architecture, query structure of DADMA is explained. Aggregate-$m$ function, for reducing the number of packets conveyed through the WSN and dilute-$m$ function for reducing the number of sensor nodes involved in a query are introduced. In the last part of the section, implementation of DADMA scheme is discussed. The figures and the pseudo code explaining our new aggregation/dilution algorithm are given.

The analytical models for the performance evaluation of the proposed scheme are provided in Section 4. Mathematical models for calculating the percentage of diluted
nodes are developed under different deployment conditions. The probability of aggregation and the probability of event detection are also formulized to determine the analytical bounds.

Simulation results are presented in Section 5, which verify the mathematical models introduced in section 4. The improvements of DADMA on the number of aggregated data packets, and diluted nodes are evaluated. The results are interpreted for changing aggregation/dilution factors, deployment conditions and other system factors.

Finally, Section 6 concludes the thesis by grouping future projections.
2 RELATED WORK

2.1 Data Querying in Sensor Networks

Queries should be resolved in the most power efficient way in WSNs. This can be achieved by reducing either the number of nodes involved in resolving a query or the number of messages generated to convey the results. There is a considerable research interest to develop efficient data querying schemes for WSNs.

The active query forwarding in sensor networks (ACQUIRE) scheme [14] aims to reduce the number of nodes involved in queries. In ACQUIRE each node that forwards a query tries to resolve it. If the node resolves the query, it does not forward it further but sends the result back. Nodes collaborate with their \( n \) hop neighbors. The parameter \( n \) is named as the look ahead parameter. If a node cannot resolve a query after collaborating with \( n \) hop neighbors, it forwards it to another neighbor. When look ahead parameter \( n \), is 1, ACQUIRE performs as flooding in the worst case.

Mobility assisted resolution of queries in large scale mobile sensor networks (MARQ) [13] makes use of the mobile nodes to collect data from the sensor network. In MARQ every node has contacts that are some of the other nodes. When contacts move around, they interact with other nodes and collect data. Nodes collaborate with their contacts to resolve the queries.

Another approach for efficient querying in sensor networks is to divide a sensor field into sub-regions, and to assign a specific number of nodes to every task set in each sub-region [15]. The number of nodes in each sub-region varies because of the non-homogenous distribution of nodes. Hence the cost of querying sensor field varies in different sub-regions. To balance this cost, forming task sets (TS) with a specific amount of nodes in each sub-region is proposed in [15]. Here, user has the initiative to trade off between accuracy/reliability and communications cost. The number of nodes in a task set indicates the resolution of the data that can be collected by querying the task set. The higher number of nodes in a task set implies higher accuracy and reliability of the system. On the other hand, more power is consumed for resolving a query as the number of nodes in a task set increases.
Another approach for querying smaller number of nodes is running decentralized spatial queries to decrease energy consumption and response time as proposed in [19]. To establish that “Peer-tree”, a hierarchical rectangular-shaped clustering method based on the number of sensor nodes in each clusters, is proposed [19]. This scheme is designed for running spatial queries efficiently, like nearest neighbor, constrained nearest neighbor or reverse nearest neighbor queries. However, assigning globally unique identifiers to each node, forming clusters and indexing nodes spatially may result in more energy consumption. Energy efficiency of “Peer-tree” method is a question mark since there are no performance evaluations in [19].

Sensor query and tasking language (SQTL) [7] is proposed as an application layer protocol that provides a scripting language. SQTL supports three types of events, which are defined by keywords receive, every, and expire. Receive keyword defines events generated by a node when the node receives a message; every keyword defines events occurred periodically; and expire keyword defines the events occurred when a timer is expired. If a node receives a message that is intended for it and contains a script, the node executes the script.

2.2 Data Aggregation in Sensor Networks

In this sub-section, some important data aggregation techniques in literature are summarized.

The scheme explained in [4] focuses on a class of aggregation predicate that is particularly well suited to the in-network regime. The exclusion problem of a sub-tree occurs when a child sensor node misses a message. This is solved by a technique called pipelined aggregation. In order to reduce the number of packets conveyed through the network, new techniques are also proposed in [4]. These techniques use standard aggregation functions (max, min, average, count and sum). However, they are not complemented with tools for spatial data aggregation.

An ad-hoc wireless network is constructed and an aggregation service is implemented in [9]. This network is perceived as a distributed database and an improved form of SQL extracts information from it. By using this scheme, in-network data aggregation improves the performance compared to the centralized techniques.
On the other hand, a general monitoring scheme for wireless sensor networks is introduced in [17] where aggregation is a must to minimize the usage of system resources. In the paper, three types of software are proposed for energy efficient monitoring: 1) Dumps, which collects detailed node states or logs per node for diagnosis. 2) Scans give abstract views of resource consumption in the network without referring to individual node. Scans indicate where should dumps be invoked. 3) Digests which specify the general characteristics of the network and indicates when should scans be invoked. Digests continuously collect aggregates of network properties, e.g. average energy of the sensor nodes, number of nodes in the network etc., in the background. Overlap problem in aggregation of digest messages is solved by constructing a routing tree and using a link quality profiling and rejection approach for non exemplary aggregation functions count, average and sum. But in [17] global IDs are used for sensor nodes which is not preferred for most WSN applications. This approach can expose serious problems when scalability is taken into consideration.

In [16] a statistical distributed estimation method is proposed for periodic aggregation. This paper focuses on fusion of sensed data by parametresizing an energy-accuracy trade-off. But network wide aggregation is not researched globally and nothing has been proposed how non-exemplary aggregation functions will be held. So in my opinion, this method can only be used for application specific problems.

The most similar approach to dilution was proposed in [33], where sensor nodes are divided in mutually exclusive sets and only one set is active at any time for energy efficiency. But in that method complex algorithms are used to determine each set for monitoring the whole sensor area. So this makes the proposed method central, static and not applicable under random deployment conditions. In contrast to [33], dilution parameters can be given in each query for different deployment conditions.

2.3 Routing in Sensor Networks

Routing in sensor networks has attracted many researchers as a new research field recently. The routing protocols can be broadly classified as data centric, hierarchical or location based [36].
Data centric routing protocols for sensor networks

Flooding, gossipping, rumor routing [37], sensor protocols for information via negotiation (SPIN)[2], sequential assignment routing (SAR) [5], directed diffusion [3], energy aware routing [38] are examples for the protocols that fall in this category. In this section we explain two of these protocols:

![SPIN protocol diagram](image)

SPIN is based on the advertisement of data available in sensor nodes. When an sensor node has a data to send, it broadcasts an advertisement (ADV) packet. The nodes interested in this data reply back by a request (REQ) packet. Then the sensor node disseminate the data to the interested nodes by using data (DATA) packets. When a node receives data, it also broadcasts an ADV, and relay DATA packets to the nodes that send REQ packets. Hence the data is delivered to every node that may have an interest. This process is shown in Figure 2.1.

In SPIN routing process is stimulated by sensor nodes. Another approach, namely directed diffusion, is sink oriented. In directed diffusion sink floods a task throughout the sensor network. While the task is being flooded, sensor nodes record the nodes which send the task to them as their gradient, and hence the alternative paths from sensor nodes to the sink is established. When there are data to send to the sink, this is forwarded to the gradients. One of the paths established is reinforced by the sink. After that point, the packets are not forwarded to all of the gradients but to the gradient in the reinforced path. Directed diffusion is illustrated in Figure 2.2.
Hierarchical routing protocols for sensor networks

Low energy adaptive clustering hierarchy (LEACH) [10], power efficient gathering in sensor information systems (PEGASIS) [39] are in this category. These techniques tackle with scalability factor by clustering nodes for routing. For example, in LEACH any sensor node can elect itself as a cluster head at any time with a certain probability. Sensor nodes access the network through the cluster head that requires minimum energy to reach.

Location based routing protocols for sensor networks

Location based algorithms such as minimum energy communication network (MECN) and small MECN (SMECN) [40] make routing decisions based on geographic locations of sensor nodes. In SMECN it is assumed that the exact locations of sensor nodes are known. Based on these locations, a sensor network is represented as a graph. Then the sub-graph that connects all nodes with minimum energy cost is computed by using a graph theoretic approach.

Figure 2.2. Directed Diffusion

a. Task dissemination  b. Gradient establishment  c. Reinforced path
3 DATA AGGREGATION AND DILUTION IN WIRELESS SENSOR NETWORKS

In this section, general architecture, query structure and implementation of our Data Aggregation and Dilution by Modulus Addressing (DADMA) scheme is discussed. DADMA is based on a simple aggregation algorithm that runs on sensor nodes. It also uses two hash functions to spatially group the sensed data.

3.1 System Architecture

In DADMA, a sensor network is considered as a distributed relational database composed of a single view that joins virtual local tables named Virtual Local Sensor Node Tables (VLSNT) located at sensor nodes. Figure 3.1 shows the distributed database perception of DADMA. In this structure an interest message, which is a query statement in our scheme, given through an external interface is disseminated to the network. Every sensor node receiving the statement firstly decides whether it is required to be involved in the query or not. If the sensor node is not diluted, it senses the given task and creates a record in its virtual local sensor node table (VLSNT). Records in VLSNT, are measurements taken upon a query arrival and consist of two fields: task and amplitude. Since a sensor node may have more than one sensor attached to it, task field indicates the sensor, e.g., temperature sensor, humidity sensor, etc., that takes the measurement. Since sensor nodes have limited memory capacities, they do not store the results of measurements. Therefore there can be a single reading for each sensor attached to a node, and task field is the key field in the VLSNT created upon a query arrival. Our perception of WSNs makes relational algebra practical to retrieve the sensed data without much memory requirement, which is different from the scheme explained in [4] where the sensed data for each task are maintained at a different column in a table.

Sensor Network Database View (SNDV) can be created temporarily either at the sink, i.e., the node that collects the data from the sensor network, or at an external proxy server. An SNDV record has three fields: task, location and amplitude. While
data are being retrieved from a sensor node, the location of the sensor node is also added to the sensed data. Since multiple sensor nodes may have the same type of sensors, i.e., multiple sensors can carry out the same sensing task, task and location fields become the key in an SNDV. In applications where nodes are not location aware, it is also possible to replace the location field with the local identifications of the reporting nodes. The location field can also be used to identify a group of nodes according to the aggregate and dilute functions explained below. It should be noted that SNDV is a temporary view where the results of a query are collected.

For many WSN applications, the sensed data are needed to be associated with the location data. For example, in target tracking and intrusion detection WSNs, sensed data are almost meaningless without relating them to a location. Therefore, location awareness of sensor nodes is a requirement imposed by many WSN applications. There are a number of practical location finding techniques for WSNs reported in [11].

Since each query results in a new SNDV, to keep the aggregated/diluted history of a WSN, it may be needed to maintain a permanent External Sensor Network Database Table (ESNDT) in a remote proxy server. In ESNDT the records obtained from queries, i.e., the records in SNDVs, are stored after being joined with a time label. For example, a daemon can generate queries at specific time intervals or at the occurrence of a specific event, and insert the records of SNDVs resulting from these queries into the ESNDT. To distinguish the equal amplitudes sensed by the same node about a specific task during different periods, task, location and time fields make the key in an ESNDT.
3.2 Querying Sensor Network in DADMA

A statement that has the structure given in Figure 3.2 starts a query. This structure is largely a part of the SQL standard [8] except for the last field starting with *based on* which will be explained later in this section. Using SQL style statements for a generic query interface have some advantages as described in [4]. Programmers and system administrators can use this practical and standard interface for all kind of WSN applications. Hardware design for WSNs can also be optimized to run this language.

In the *Select* keyword of the SQL statement common aggregation functions such as *avg*, *min*, *max* can be used to indicate how to aggregate the amplitude field. The fields to be projected from an ESNDT are also listed after this keyword. *From* keyword indicates the nodes to be involved in the query. *Any* means that even a single node may be enough to resolve the query, and any node in the sensor network can do it. When *every* keyword is used, all of the nodes in the sensor network are supposed to be involved in the query. When a *task* or a set of tasks is given in the query, only the sensors in the specified types carry out the measurement. We also introduce *aggregate* and *dilute* keywords to spatially group the nodes. *Where* keyword is for defining selection conditions according to available power and/or time and/or amplitude and/or location. *Group by* field is used to specify the set of tasks for which the aggregation of the sensed data will be carried out. *Based on* keyword is followed by the parameters required for the aggregation and dilution algorithm run by the sensor nodes.
Select [task, time, location, [distinct | all], amplitude,
[[avg | min | max | count | sum ] (amplitude)]
From [any, every, task, aggregate-m, dilute-m]
Where [power available [<>] PA]
   Location [in | not in] RECT |
   t_{min}<time<t_{max} |
   amplitude [<>===>] a]
Group by task
Based on [time limit = l, | packet limiy = l_p |
   resolution =r |   region = xy]

Figure 3.2. The structure of an SQL statement for DADMA.

A user can retrieve a subset of data fields available in an SNDV, and can aggregate the amplitude field either by grouping data based on task and/or by using aggregate-m function given in Equation 3.1. Some of the sensor nodes can also be excluded from a query by dilute-m function in Equation 3.2.

\[ f(x) = x \div m \] (3.1)
\[ f(x) = \left( \frac{x}{r} \right) \mod \left( \frac{m}{r} \right) \] (3.2)

where

\( x \) is the grid location of a node relative to one of the axes,
\( r \) is the resolution in meters, and
\( m \) is the dilution or aggregation factor.
When dilute-\(m\) command is given by the user, every sensor first uses Equation 3.2 to find its location indices in horizontal and vertical axes and then compare the indices with the region values \(x\) and \(y\) sent in the “based on” field of the query. If they match, the sensor node replies the query. For example, the location indices of a sensor node at location \{46, 74\} are \{3, 1\} for \(m=8\) and \(r=2\). Therefore, if the region value in the query is \{3, 1\}, this sensor should make measurement. Hence, only the sensor nodes in \(r \times r\) meter squares located in every \(m\) meters respond to the query as shown in Figure 3.3, and the others stay idle. This is a practical node dilution technique especially when sensor nodes are randomly deployed according to uniform distribution, and the sensor network is monitoring environmental conditions such as temperature, humidity and pressure.

For the same example the indices found out by using Equation 3.1 are \{5, 9\}. When aggregate-\(m\) command is received, the values measured by a sensor node are aggregated with the values measured by the other nodes having the same indices. Hence, we can address the sensor nodes at certain geographic locations, and aggregate the sensed data in the same rectangular region as shown in Figure 3.4.
3.3 Implementation of DADMA

Users or the system administrators prepare an SQL statement in order to query a WSN. This query is disseminated in the entire network. When a sensor node involved in the task dissemination process receives a query, it first repeats the query. The dissemination of the query is performed by the routing algorithm which works independent from DADMA. Then the node sends back a reply packet to the node that it receives the query packet from, i.e., the gradient. Hence gradients know that there are some children nodes that will send them data packets. The sensor node receiving a query packet checks the conditions specified in the from field of the SQL statement. If the query includes a dilute-m function and the location of the sensor is not equal to the region value in the based on field according to the dilute-m, the sensor is diluted. Thus, the sensor node is out of sensing process. If the sensor is not diluted, then it checks whether the power, time and location conditions in the SQL statement are met. When all these conditions are satisfied the sensor node carries out the measurements and inserts the results into its VLSNT.
if (newtaskreceived(query)) {
    t = currenttime;
    broadcast(query); // repeat the received query
    send("reply")
    received = 0;
    notleaf = 0;
    localSNDV = createLocalSNDV();
    VLSNT = createVLSNT();
}

if (taskavailable(query.task)) {
    run (query.task, amplitude);
    addtoVLSNT (query.task, amplitude, VLSNT);
}

while (!notleaf && currenttime - t < l_min) {
    if (newdatareceived (data)) {
        if (data.type == "reply")
            notleaf++;
        else{
            received++;
            addtolocalSNDV(data, localSNDV);}
    }
}

if (!notleaf) { // this is a leaf node
    if (!NotEmpty(VLSNT)) {
        makeDataPacket(dataPacket, VLSNT);
        send(dataPacket);
        send("push");
    }
    else {
        while (!notleaf && currenttime - t < l_i) {

            if (newdatareceived(data)) {
                if (data.type == "reply")
                    notleaf++;
                else if (data.type == "push")
                    notleaf--;
                else{
                    received++;
                    addtolocalSNDV(data, localSNDV);}
            }

            if (received >= l_s) { // buffer size is reached
                aggregate(dataPacket, localSNDV, VLSNT);
                send(dataPacket);
                initializeTables(localSNDV, VLSNT);
            }
        }
        if (!NotEmpty(VLSNT) || notEmpty(localSNDV)) {
            aggregate(dataPacket, localSNDV, VLSNT);
            send(dataPacket);
            send("push");
        } // end else
    }
}

Figure 3.5. The DADMA Algorithm.
After relaying a query, sending node also starts a timer. If it expires before receiving a reply packet during \(l_{\text{min}}\), which is derived from \(l\) specified in the SQL statement, the node understands that it will not receive any data packet from its neighbors. In such case, it sends the related records from the VLSNT to its gradient, and then a push packet meaning that it has no more to deliver.

If the sensor receives a reply packet before the \(l_{\text{min}}\) time expires, it understands that there are children nodes, and it waits to receive data from them. The sensor will await the sensed data from every node that sends a reply packet. However, certain failures may occur in the child and/or grandchild nodes due to energy lack or some other external reasons after a node sends a reply packet. In order to handle this situation, all nodes run an \(l\) timer. When \(l\) expires, the node terminates the process for the query. Therefore, if nodes that sent reply packet cannot send push packet in \(l\) period, then their transmission is neglected.

Nodes do not buffer more than \(l_p\) packets during the \(l\) period. When \(l_p\) packets are received, they are aggregated and sent to the gradient. If there is an aggregate-\(m\) function in the SQL statement, data packets should be grouped by task and location and be aggregated accordingly. Thus, the data sensed by the nodes in each rectangle of the virtually partitioned WSN become aggregated. In the absence of an aggregate-\(m\) function, it will suffice to group the sensed data only by task. If the aggregation function specified in the select keyword of the SQL like statement is average, for the sake of the consistency of the distributed aggregation both count and sum should be calculated and sent to the gradient. After sending the aggregated data, the node notifies its gradient that it has no more to deliver by sending a push packet. The algorithm for this procedure is given in Figure 3.5.

In this procedure \(l\) and \(l_p\) are important parameters given in the SQL statement. The other parameter \(l_{\text{min}}\) is derived from \(l\), i.e., \(l_{\text{min}}=a \times l\), where \(0<a\leq1\). These parameters, especially \(l_p\), can have an impact on the performance of the algorithm because the number of data packets that can be stored in sensor nodes is also limited by the memory space available in nodes. The sensitivity of DADMA against \(l_p\) is examined in detail in our experiments.

An illustrative example is given in Figure 3.6 where the procedure is applied to a six node WSN. In the first step shown in Figure 3.6, Node 1 disseminates a query,
the query is received by two children nodes $b$ and $c$. As soon as $b$ and $c$ receive the query, they repeat it and send back a reply packet in Step 2. The query broadcasted by $b$ is received by two more nodes $d$ and $e$, which also repeat the query and send back a reply packet in Step 3. However, $c$ does not have any children node that can hear it. Therefore it waits during $l_{\min}$, and understands that it is a leaf node at the end of this period because it does not receive a reply packet. Since there is no node that will send data to Node $c$, it first sends back the data in its VLSNT to Node $a$, and then a push packet as shown in Step 4. When Node $a$ receives a push packet from Node $c$, it understands that no more data will be received from Node $c$. Node $e$ and $f$ follow a similar procedure. When a node receives push packets, which are equal to the reply packets in number, it understands that all its children have reported, and therefore it also reports to its parent and then sends a push packet. Based on this, Node $b$ first sends the collected data and then a push packet to Node $a$ when it receives push packets from both Nodes $d$ and $e$ as shown in Step 10.
Figure 3.6. An illustrative example for the implementation of DADMA.
4 PERFORMANCE EVALUATION

In the previous sections, aggregate-m function, for reducing the number of packets conveyed through the WSN and dilute-m function for reducing the number of sensor nodes involved in a query are introduced. In this section the mathematical models for dilution and aggregation functions will be developed to give the performance evaluation of our new scheme DADMA. Since dilution decreases the number of sensing nodes, formulating the percentage of diluted nodes is important to configure the algorithms utilization. On the other hand, dilution has an effect on detection of events. If some of the sensor nodes will be out of sensing task, this may cause undetected events in the case of dilution. So the tradeoff between the number of diluted nodes and event detection is formulated to determine the analytical bounds. Finally, the probability that aggregation takes place is mathematically given to criticize the performance gains.

Figure 4.1. Sensor field
The model shown in Figure 4.1 is used to define the parameters such as the dilution factor $m$, and the dilution resolution $r$. In this model the smallest rectangle that confines all sensor nodes of a sensor network represents a sensor field. The width of the sensor field is $w$, and the height of it is $h$. The sensor field is considered to be covered with fixed sized square grids with length $r$. The sensor nodes are randomly deployed according to a given distribution e.g., uniform, exponential, gaussian, etc, in the sensor field.

4.1 The Percentage of Diluted Nodes

According to the distribution function and the number of nodes deployed, every virtual cell in Figure 4.1 contains a number of sensor nodes. In that figure shaded cells represent the dilution cells. Positions of the sensor nodes can be defined by a coordinate system based on a corner of the sensor field, and will be represented as $(x_i, y_i)$ for the $i$'th sensor. Based on this model, the probability, $P_{kl}$, that the node $a(x,y)$ is in the virtual cell at column $k$ and row $l$ is

$$P_{kl} = P(k \times r < x_a \leq (k+1) \times r, l \times r < y_a \leq (l+1) \times r) \quad (4.1)$$

It is easy to see that $X$ and $Y$ are independent random variables. So from the basic probability theory Equation 4.1 can be written as

$$P_a = (F_{Xa}((k+1)\times r) - F_{Xa}(k\times r))(F_{Ya}((l+1)\times r) - F_{Ya}(l\times r)) \quad (4.2)$$

where $F_{Xa}$ and $F_{Ya}$ are the probability distribution functions (pdf) of $x_a$ and $y_a$ respectively. For dilution, there is one non-diluted cell in every $m \times m$ squares. So the total number of non-dilution cells, $N_c$, is given by

$$N_c = \frac{w}{m} \times \frac{h}{m} \quad (4.3)$$

where $w$ is the width and $h$ is the height of the sensor field.

By generalizing Equation 4.2 for the whole sensor field where $\delta$ is the index of the first non-dilution cell, we find

$$P_a = 1 - \left( \sum_{k=\delta \frac{m}{r} \times \frac{2m}{r}, \frac{w-r}{m} \times \frac{m}{r}} F_{Xa}((k+1) \times r) - F_{Xa}(k \times r) \right)$$
\[
\times \sum_{l=0}^{m-2} \frac{m}{m-1} \frac{m}{r} F_{1}(l+1 \times r) - F_{1}(l \times r) \quad (4.4)
\]

Percentage of Diluted Nodes = \( P_{a} \times 100 \quad (4.5) \)

Number of Diluted Nodes \( = P_{a} \times N \quad (4.6) \)

where \( N \) is the number of sensor nodes deployed, and \( P_{a} \) is the probability of dilution. When we assume that the nodes are randomly deployed according to uniform distribution, probability that a node is diluted is given by

\[
P_{a} = 1 - \left( \int_{0}^{w} dx \int_{0}^{r} dy \right) - \left( \int_{0}^{h} dx \int_{0}^{r} dy \right) + \left( \int_{0}^{w} dx \int_{0}^{h} dy \right)
\]

\[
= 1 - \left( \frac{m}{w} \times \frac{m}{r} \right) = 1 - \frac{r^2}{m^2} \quad (4.7)
\]

where \( X \) and \( Y \) are independent, uniform random variables in \((0,w)\) and \((0,h)\) respectively. For uniform distribution it can be easily seen that only dilution factor \( m \) and dilution resolution \( r \), determine the percentage of diluted nodes. Neither the size of the sensor field nor the number of sensor nodes has any effect on dilution. This is also intuitively clear which proves Equation 4.4. Please note that Equation 4.4 can be extended for other distributions. In the next section the simulation results will figure out if there are other parameters affecting the percentage of dilution for other deployment distributions.

4.2 Probability of Detection for Dilution Factor is \( m \) and Dilution Resolution \( r \)

When dilute-m function is used, a number of sensor nodes will be excluded from the sensing task. The percentage of diluted nodes is given in Equation 4.5 and the number of diluted nodes is given in Equation 4.6. Dilution has an impact on the probability that an event is detected by a sensor network. Since detection of an event is determined by the sensing range of nodes and the distance between the event and the nodes, the probability that an event can be detected by at least one sensor \( P_{\beta} \), for dilution factor \( m \) and dilution resolution \( r \), can be formulated as:
\[ P_\phi = 1 - P_{\phi}^N \quad (4.8) \]
\[ P_{\phi} = 1 - (P_{\phi} \times (1 - P_\phi)) \quad (4.9) \]

where \( N \) is the number of sensor nodes, \( P_\phi \) is the probability that a sensor node cannot detect the event, \( P_\phi \) is the probability that the event is in the sensing range of a node, \( P_\phi \), given in Equation 4.4, is the probability that a node is diluted.

\[ P_\phi = P\left( \sqrt{(x_n - x_e)^2 + (y_n - y_e)^2} < d_n + d_e \right) \quad (4.10) \]

where \((x_n, y_n)\) is the coordinate of the sensor node, \((x_e, y_e)\) is the coordinate of the event to be detected, \( d_n \) is the sensing radius of the sensor node and \( d_e \) is the effect radius of the event. We find \( P_\phi \) in two steps. First, we compute the probability density functions (pdf) of \( X = X_e - X_n \) and \( Y = Y_e - Y_n \); then we compute the pdf of \( Z = \sqrt{(X_e + Y)^2} \).

At the first step by substituting \( X_n = X + X_n \) and \( Y_n = Y + Y_n \), we find

\[ f_X(x) = \int_{-\infty}^{\infty} f_{X_e}(x + x_e)dx_e \quad (4.11) \]
\[ f_Y(y) = \int_{-\infty}^{\infty} f_{Y_e}(y + y_e)dy_e \quad (4.12) \]

Since \( X_n \) and \( X_e \), \( Y_n \) and \( Y_e \) are independent random variables,

\[ f_X(x) = \int_{-\infty}^{\infty} f_{X_e}(x + x_e)f_{X_e}(x_e)dx_e \]
\[ f_Y(y) = \int_{-\infty}^{\infty} f_{Y_e}(y + y_e)f_{Y_e}(y_e)dy_e \quad (4.13) \]

At the second step to formulate the pdf of \( Z \), an auxiliary random variable \( T \), as \( T = X \), is introduced. This will enable us to use the general formula of finding \( f_{zt} \) from two functions of two random variables with \( n \) real roots, given below

\[ f_{zt}(z, t) = \sum_{i=1}^{n} f_{x,y}(x_i, y_i) \quad (4.14) \]
The equations
\[ Z - \sqrt{X^2 + Y^2} = 0 \] \[ T - X = 0 \] (4.15)

have two real roots, for \(|t| < z\), namely
\[
x_1 = t \quad x_2 = t
\]
\[
y_1 = \sqrt{z^2 - t^2} \quad y_2 = -\sqrt{z^2 - t^2} \] (4.16)

At both roots, \(|\|\) has the same value:
\[
|J_1| = |J_2| = \frac{z}{\sqrt{z^2 - t^2}} \] (4.17)

Since X and Y are independent random variables, a direct application of Equation 4.14 yields
\[
f_{zt}(z,t) = \frac{z}{\sqrt{z^2 - t^2}} \left[ f_X(x_t)f_Y(y_t) + f_X(x_t)f_Y(y_t) \right] \] (4.18)

We get \(f_X(x)\) and \(f_Y(y)\) in Equation 4.13, so we can find \(F_Z(d_n + d_e)\)
\[
f_Z(z) = \int_{-\infty}^{z} f_{zt}(z,t) dt \] (4.19)
\[
P_\phi = F_Z(D_n + d_e) = \int_{0}^{d_n + d_e} f_Z(z)dz \] (4.20)

where \(P_\phi\) is the probability of detection.

Equation 4.20 that gives probability of detection can be extended for the Gaussian distributions, where \(X_n, X_e, Y_n\) and \(Y_e\) are distributed according to \(N(0, \sigma_{x_n}^2)\), \(N(0, \sigma_{x_e}^2)\), \(N(0, \sigma_{x_n}^2)\), \(N(0, \sigma_{x_e}^2)\) respectively, under \(\sigma_{x_n} - \sigma_{x_e} = \sigma_{y_n} - \sigma_{y_e}\) condition. If we substitute functions in Equation 4.18, we get
\[
f_{zt}(z,t) = \begin{cases} 
\frac{1}{\pi \sigma^2} \frac{z}{\sqrt{z^2 - t^2}}, & z > 0, |t| < z \\
0, & \text{otherwise}
\end{cases} \] (4.21)
\[ f_z(z) = \int_{-\infty}^{\infty} f_{z|x}(z,t) dt \]

\[ = \frac{z}{\sigma^2} e^{-z^2/2\sigma^2} \left[ \frac{2}{\pi} \int_0^\infty \frac{dt}{\sqrt{z^2 - t^2}} \right] u(z) \quad (4.22) \]

Since the term in parenthesis has value \( \pi/2 \), \( Z = \sqrt{X^2 + Y^2} \), \( f_z(z) \) is the Rayleigh density function where standard deviation \( \sigma \) is,

\[ \sigma = \sigma_x - \sigma_y = \sigma_{y_x} - \sigma_{y_y} \quad (4.23) \]

\[ P_\theta = F_z(d_x + d_z) = \int_0^z \frac{z}{\sigma^2} e^{-z^2/2\sigma^2} d(z) \quad (4.24) \]

Equation 4.20 will also be extended for uniform random variables \( X_\delta(0,w) \), \( X_\delta(0,w) \), \( Y_\delta(0,h) \) and \( Y_\delta(0,h) \) where \( w \) and \( h \) are the width and height of the sensor field respectively.

If we solve Equation 4.13 for these random variables, we get

\[ f_x(x) = \begin{cases} \int_0^{+\infty} \frac{1}{w^2} d x = \frac{w+x}{w^2}, & -w \leq x \leq 0 \\ \int_0^{-\infty} \frac{1}{w^2} d x = \frac{w-x}{w^2}, & 0 \leq x \leq w \end{cases} \quad (4.25) \]

\[ f_y(y) = \begin{cases} \int_0^{+\infty} \frac{1}{h^2} d y = \frac{h+y}{h^2}, & -h \leq y \leq 0 \\ \int_0^{-\infty} \frac{1}{h^2} d y = \frac{h-y}{h^2}, & 0 \leq y \leq h \end{cases} \quad (4.26) \]

Same steps are followed from Equation 4.15 to Equation 4.17 and then \( f_{x|x_1}, f_{x|x_2}, f_{y|y_1} \) and \( f_{y|y_2} \) are substituted in Equation 4.18.

\[ f_{z|x}(z,t) = \begin{cases} \frac{w+t}{w} \frac{h+\sqrt{z^2-t^2}}{h} + \frac{w-t}{w} \frac{h-\sqrt{z^2-t^2}}{h}, & -w \leq x \leq 0, -h \leq y \leq 0 \\ \frac{w+t}{w} \frac{h+\sqrt{z^2-t^2}}{h} + \frac{w-t}{w} \frac{h+\sqrt{z^2-t^2}}{h}, & -w \leq x \leq 0, 0 \leq y \leq h \\ \frac{w+t}{w} \frac{h-\sqrt{z^2-t^2}}{h} + \frac{w-t}{w} \frac{h+\sqrt{z^2-t^2}}{h}, & 0 \leq x \leq w, -h \leq y \leq 0 \\ \frac{w+t}{w} \frac{h-\sqrt{z^2-t^2}}{h} + \frac{w-t}{w} \frac{h-\sqrt{z^2-t^2}}{h}, & 0 \leq x \leq w, 0 \leq y \leq h \end{cases} \]

\[ f_{z|t}(z,t) = \frac{z}{\sqrt{z^2-t^2}} \begin{cases} \frac{2}{h^2} \left( \frac{1+t}{w^2} \right), & -w \leq x \leq 0, -h \leq y \leq h \\ \frac{2}{h^2} \left( \frac{1-t}{w^2} \right), & 0 \leq x \leq w, -h \leq y \leq h \end{cases} \quad (4.27) \]
where \( z > 0, |y| < z \) conditions must be satisfied.

\[
f_Z(z) = \int_{-\infty}^{\infty} f_{ZI}(z, t) dt
\]

\[
f(z) = \left\{ \begin{array}{ll}
\frac{2z}{h^2} \arcsin \frac{t}{\sqrt{z^2 - t^2}} + \frac{2z}{h^2} \arcsin \frac{\sqrt{z^2 - t^2}}{z}, & -w \leq x \leq 0, -h \leq y \leq h \\
\frac{2z}{h^2} \arcsin \frac{t}{\sqrt{z^2 - t^2}} - \frac{2z}{h^2} \arcsin \frac{\sqrt{z^2 - t^2}}{z}, & -w \leq x \leq 0, -h \leq y \leq h
\end{array} \right. \tag{4.28}
\]

if we substitute \( v = z^2 - t^2 \), so \( dv = -2w dw \) and solve the integrals, since \( z \geq 0 \) and \( |w| < z \) we get:

\[
f_Z(z) = \left\{ \begin{array}{ll}
\frac{2z}{h^2} \arcsin \frac{\frac{\pi}{2} + \frac{z}{w}}{\sqrt{z^2 - \frac{\pi}{2} - \frac{z}{w}}}, & -w \leq x \leq 0, -h \leq y \leq h \\
\frac{2z}{h^2} \arcsin \frac{\frac{\pi}{2} - \frac{z}{w}}{\sqrt{z^2 - \frac{\pi}{2} + \frac{z}{w}}}, & -w \leq x \leq 0, -h \leq y \leq h
\end{array} \right. \tag{4.29}
\]

The probability that the event is in the sensing range of a node becomes:

\[
P_y = F_Z(d_u + d_v) = \int_{0}^{d_u + d_v} f_Z(z) dz
\]

\[
P_y = \left\{ \begin{array}{ll}
\frac{\pi (d_u + d_v)^3}{2h^2} + \frac{2(d_u + d_v)^3}{3h^2}, & -w \leq x \leq 0, -h \leq y \leq h \\
\frac{\pi (d_u + d_v)^3}{2h^2} - \frac{2(d_u + d_v)^3}{3h^2}, & 0 \leq x \leq w, -h \leq y \leq h
\end{array} \right. \tag{4.30}
\]
4.3 The Probability of Aggregation When *Aggregate*-m Function is Used

![Figure 4.2: Sensor field when aggregate m function is used.](image)

When the aggregate-m function is used, the sensor field is divided into equal squares with size mxm as in Figure 4.2. Each sensor node aggregates its sensed data with the data from the nodes in the same virtual cell. So the probability that a data packet is aggregated with another packet while being conveyed in the network is equal to the probability that an event can be detected by at least two sensor nodes in the same cell.

\[ P_{\lambda} = 1 - (P_{\Phi} + P_{\tau}) \]  \hspace{1cm} (4.31)

where \( P_{\lambda} \) is the probability of aggregation, \( P_{\Phi} \) is the probability that none of the events will be detected, \( P_{\tau} \) is the probability that there will be one detection at least in one aggregation cell. Since \( P_{\Phi} \) is the probability of detection given in Equation 4.20 and \( N \) is the number of sensor nodes in the WSN

\[ P_{\Phi} = (1 - P_{\Phi})^{N} \]  \hspace{1cm} (4.32)

\[ P_{\tau} = P_{\Phi} + P_{2\Phi} \sum_{i=0}^{\text{wcell}} (P_{i} \times \sum_{k=i+1}^{\text{wcell}} P_{k}) \]

\[ + P_{3\Phi} \sum_{i=0}^{\text{wcell}} (P_{i} \times \sum_{k=i+1}^{\text{wcell}} (P_{k} \times \sum_{l=k+1}^{\text{wcell}} P_{l})) + ... + P_{n\Phi} \sum_{i=0}^{\text{wcell}} (P_{i} \times \sum_{k=i+1}^{\text{wcell}} P_{k} \times \sum_{l=k+1}^{\text{wcell}} P_{l}) \]  \hspace{1cm} (4.33)
where

\[ P_{1\phi} = \binom{N}{1} \times P_{\phi}^1 \times (1 - P_{\phi})^{N-1} \]

\[ P_{2\phi} = \binom{N}{2} \times P_{\phi}^2 \times (1 - P_{\phi})^{N-2} \]

\[ P_{wh}^{\text{wm}} = \left( \frac{N}{m \times m} \right) \times P_{wh}^{\text{wm}} \times (1 - P_{\phi})^{N-\text{wm}} \] (4.34)

and \( P_i \) is the probability that the node \( a(x,y) \) is in the aggregation cell at column \( k \) and row \( l \) where

\[ k = i \mod \frac{w}{m} \] (4.36)

\[ l = \frac{i - k}{w} \times \frac{m}{w} \] (4.36)

\[ P_i = (F_{X_a}((k+1) \times m) - F_{X_a}(k \times m)) \times (F_{Y_a}((l+1) \times m) - F_{Y_a}(l \times m)) \] (4.37)

When we substitute equations 4.32, 4.33, 4.34, 4.35, 4.36 and 4.37 in equation 4.31, we get directly the \( P_a \) which is the probability of aggregation.
5 EXPERIMENTAL RESULTS

In this section, simulation results are presented which verify the mathematical models introduced in the previous section. The gains of DADMA in the number of aggregated data packets, and diluted nodes are evaluated. The impact of changing aggregation/dilution factor $m$, resolution $r$, packet limit $l_p$, the sum of the sensing radius of nodes and the effect radius of events $d_n+d_e$ are examined for various number of nodes deployed over an area $20\times20$ in size according to both the Uniform and Gaussian distributions. In simulations, it is assumed that every node has data to respond the queries in order to ensure the fairness and to test our system against the worst-case scenario. Since physical layer parameters, such as path loss exponent and hop distance, do not affect our results, they are not factoring parameters in our experiments.

Simulations are performed using MATLAB 6.5. First all of the mathematical models formulated in the previous section are coded in Matlab and the results are given in related graphics for different factoring variables. At the second step a simple wireless sensor network is modeled in Matlab and Directed Diffusion [3] is implemented as the routing algorithm. In that scenario, number of aggregated packets is simulated for different deployment distributions and for different network properties.

In Figures 5.1, 5.2 and 5.3 the performance of the system is evaluated based on the percentage of nodes involved in a query when dilution is used, and nodes are randomly deployed according to the Uniform distribution. The probability of dilution for Uniform distribution is given in equation (4.7). As it can be seen from the related figures, the percentage of diluted nodes is only proportional with $r^2/m^2$. It is observed that percentage of diluted nodes increases exponentially when dilution factor $m$ increases. On the other hand the percentage of diluted nodes is higher for the smaller dilution resolution $r$, and the relation between the percentage of diluted nodes and the dilution resolution is almost linear. This is an expected result because the area covered by the nodes involved in queries enlarges as the dilution factor $m$ decreases and the dilution resolution $r$ increases. When $r$ is equal to $m$ whole area is covered,
so no sensor node is diluted. On the other hand when \( r = m/2 \), 75% of the sensor nodes are diluted.

Figure 5.1. Percentage of diluted nodes for varying dilution factors and dilution resolutions.

Figure 5.2. Percentage of diluted nodes for varying dilution factors when dilution resolution is 1 m.
Figure 5.3. Percentage of diluted nodes for varying dilution resolutions when dilution factor is 12 m.

In Figures 5.4, 5.5 and 5.6 the same experiments are carried out for the case where nodes are deployed according to the Gaussian distribution. The relation between the percentage of diluted nodes and the dilution factor $m$ and resolution $r$ parameters are almost the same as the one found for the Uniform distribution. Therefore, we can say that the deployment of nodes do not have an important impact on the percentage of diluted nodes when $dilate-m$ function is used.
Figure 5.4. Percentage of diluted nodes for varying dilution factors and dilution resolutions.

Figure 5.5. Percentage of diluted nodes for varying dilution factors when dilution resolution is 1 m.
In Figures 5.7, 5.8 and 5.9, the probability that an event can be detected by at least one sensor is shown for varying dilution factor $m$ and dilution resolution $r$. There are 20 sensor nodes deployed according to Gaussian distribution with standard deviation 4, and the sum of the sensor node sensing range and the effect radius of an event $d_n+d_e$ is 2. In Figure 5.7, the effects of both dilution factor $m$ and dilution resolution $r$ to the probability of detection are examined at the same. It is easy to see that the part of the figure parallel to the x-y plane covers the $r$ and $m$ values, except $r$ is equal to $m$, where dilution of nodes does not decrease the probability that detection of events.

In Figure 5.8 it is observed that the increase in dilution factor $m$ also decreases the probability of detection. However in Figure 5.9, the increase in dilution resolution $r$ increases the probability of detection. Since the event size and the detection range is fixed in these experiments, up to a certain dilution resolution and factor, the probability of detection is same with the value when no dilution is used, because the event size and the detection range covers an area larger than a cell of dilution, i.e., an $r\times r$ square, and therefore the event falls in the coverage area of at least one set of not diluted cells. This relation can easily be seen in Figure 4.1.
Figure 5.7. Probability of detection for varying dilution factors and dilution resolutions.

Figure 5.8. Probability of detection for varying dilution factors when dilution resolution is 2 m.
Figure 5.9. Probability of detection for varying dilution resolutions when dilution factor is 8 m.

Figure 5.10. Probability of detection for varying number of sensor nodes and dilution factors when dilution resolution is 3m, $d_n+d_e=2$.

In Figure 5.10, the probability that an event can be detected by at least one sensor is shown for varying dilution factor $m$ and the number of sensor nodes between 2 to 20. Other parameters remain same with Figures 5.7, 5.8 and 5.9. We also show the case where no dilution is applied in this figure. DADMA provides a tradeoff mechanism
between the cost of queries, i.e., message complexity, and accuracy/reliability. This figure is to provide a better insight about this tradeoff. For the lower dilution factor, the difference between the accuracy of non-diluted and diluted cases is minimal as shown in Figure 5.10.

In Figures 5.11 and 5.12, the probability that an event can be detected by at least two sensor nodes in the same cell, which is the probability of aggregation, is simulated for varying aggregation factor $m$ and number of sensor nodes respectively. In both simulations sensor nodes are deployed according to Gaussian distribution with standard deviation 4 and the sensing range of a sensor node plus the effect radius of an event $d_n + d_e$ is 2. It is obvious that the probability of aggregation is very high for all $m$ values and benefits of this highly probable aggregation will be shown graphically in the next simulations.

![Figure 5.11. Probability of aggregation for varying aggregation factors when 64 sensor nodes are deployed and $d_n + d_e = 2m$.](image)
Figure 5.12. Probability of aggregation for varying number of sensor nodes when aggregation factor is 8 and \( dn+de=2m \).

In Figures 5.13 to 5.20 the number of the data packets transmitted while the results of a query are being conveyed through the network is shown in the case of aggregation. The simulations are run for varying aggregation factor \( m \) and the number of sensor nodes. The network layer protocol used in our simulations is directed diffusion [3] with only one gradient, which means that only one path is established to the sink for each node. Interest messages are not taken into count to show the true advantages of our spatial aggregation. Sensor nodes are deployed according to Gaussian distribution in Figures 5.13, 5.14, 5.15, 5.16, and they are deployed according to Uniform distribution in Figures 5.17, 5.18, 5.19, 5.20. In our scenario, an event is detected by all sensor nodes to eliminate the effect of event location to the simulations. It can be easily seen from our simulations that there are two factors affecting the utilization of aggregation except for aggregation factor \( m \). First one is the node density and the other one is the communication range of the sensor nodes (\( d_n \)). If the hop counts of the sensing nodes increases the number of aggregated packets also increases. On the other hand if aggregation factor \( m \) increases, the area of the aggregation cells increase and the number of aggregation cells decrease. Since the network area is fixed sized, less number of more populated sensor node groups will cause more aggregated packets for increasing \( m \) values.
Figure 5.13. Number of transmitted packets for varying aggregation factors and number of sensor nodes. Communication range of the sensor nodes $d_n$ is 2m.

Figure 5.14. Number of transmitted packets for varying aggregation factors and number of sensor nodes. Communication range of the sensor nodes $d_n$ is 3m.
Figure 5.15. Number of transmitted packets for varying aggregation factors and number of sensor nodes. Communication range of the sensor nodes $d_n$ is 4m.

Figure 5.16. Number of transmitted packets for varying aggregation factors and number of sensor nodes. Communication range of the sensor nodes $d_n$ is 6m.
Figure 5.17. Number of transmitted packets for varying aggregation factors and number of sensor nodes. Communication range of the sensor nodes $d_n$ is 2m.

Figure 5.18. Number of transmitted packets for varying aggregation factors and number of sensor nodes. Communication range of the sensor nodes $d_n$ is 3m.
In Figures 5.19 to 5.28, the number of the data packets transmitted while the results of a query are being conveyed through the network is shown when packet limit $l_p$ is used. In a real application packet limit can be set to prevent the overflow of the
sensor’s buffers. The simulations are run for varying $l_p$ and the number of sensor nodes when aggregation factor is 8. The network layer protocol used in our simulations is again directed diffusion [3] with only one gradient and sensor nodes are deployed according to Gaussian distribution in Figures 5.21, 5.22, 5.23, 5.24 and according to Uniform distribution in Figures 5.25, 5.26, 5.27, 5.28. When $l_p$ is 1, the system can not use DADMA, therefore it is the directed diffusion without DADMA case. So simulations clearly prove that aggregation decreases the number of transmitted data packets even for small $l_p$ values.

![Figure 5.21. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, dn=2)](image-url)
Figure 5.22. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, \(dn=3\))

Figure 5.23. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, \(dn=4\))
Figure 5.24. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, \(dn=6\))

Figure 5.25. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, \(dn=2\))
Figure 5.26. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, dn=3)

Figure 5.27. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, dn=4)
Figure 5.28. Number of transmitted packets for varying number of packet limits and number of sensor nodes (aggregation factor=8, \(dn=6\))
6 CONCLUSION AND FUTURE WORK

According to the characteristics of WSNs, power efficiency is one of the critical design factors all type of applications. In literature, it is showed that communication is much power consuming than processing [18]. At this point, local processing and data aggregation gains great importance as a particularly useful paradigm for routing in wireless sensor networks. Similar to aggregation, data querying is considerable since there are a sheer number of nodes with stringent energy constraints, so it may not be feasible to fetch every reading of sensor nodes for central processing.

In this thesis, a new distributed spatial data aggregation and dilution scheme based on hashing and relational algebra is introduced for wireless sensor networks. In our work, a sensor network is perceived as a distributed relational database, which is created by joining records distributed in virtual local sensor node tables located at sensor nodes. These records are collected by neighboring nodes in a sensor network database view structure and passes through an external proxy server to form the external sensor network database table. It is useful to emphasize note that sensed data are not stored at sensor nodes; they are passed through the proxy server and recorded in external sensor network database table.

Query model of DADMA uses a SQL like language which satisfies a standard interface to the operators and can be optimized for the hardware specifications. Each sensor node that receives the query runs a simple algorithm to perform the task by following the rules included in the query while conveying the data packets coming from the other nodes. The aggregation and/or dilution parameters are provided in the query. Nodes find out if they are diluted, i.e., they are not supposed to get involved in a query, by using a simple hash function. Dilution provides a tradeoff mechanism between the accuracy/reliability of a query result and the cost of the query. The data are also aggregated according to their location in the relaying nodes by using another hash function. In our scheme, the cost of queries, i.e., message complexity, is reduced, and this is achieved by running a simple distributed algorithm that can fit in the limited computation and memory capacity of tiny sensor nodes.
Three mathematical models are developed to evaluate the percentage of diluted nodes, the probability of detection for a specific dilution function and the probability of aggregation for aggregation factor $m$. According to these models, our experiments are performed using MATLAB 6.5. The gains of DADMA in the number of aggregated data packets, and diluted nodes are evaluated. The impact of changing aggregation/dilution factor $m$, resolution $r$, packet limit $l_p$, the sum of the sensing radius of nodes and the effect radius of events $d_n+d_e$ are examined for various number of nodes under different deployment conditions.

Results of our experiments show that percentage of diluted nodes increases exponentially when dilution factor $m$ increases. On the other hand the percentage of diluted nodes is higher for the smaller dilution resolution $r$, and the relation between the percentage of diluted nodes and the dilution resolution is almost linear. 75% of sensor nodes are diluted when the dilution factor is twice as much as the dilution resolution.

According to another set of experiments, it is observed that the increase in dilution factor $m$ decreases the probability of detection. However, since the event size and the detection range is fixed in the experiments, up to a certain dilution resolution and factor, the probability of detection is same with the value when no dilution is used, because the event size and the detection range covers an area larger than a cell of dilution.

On the other, hand without dilution, aggregation reduces 50% of data transmissions on the average, and manages to aggregate up to 90 to 99 % of data packets while they are being conveyed through the sensor network. These results prove the applicability and efficiency of DADMA.

There are a number of areas for future work. First, the performance of time out mechanisms is not explored. More simulations must be generated for discovering these parameters. Second, the effects of mobility and node failures are not researched. Also more aggregation and dilution functions can be developed.
REFERENCES


Data Communications in Latin America and the Caribbean, Costa Rica, April, 2001.


BIOGRAPHY

Tolga Çöplü was born in Çorum in 1979. He graduated from Çorum Anatolian High School in 1997 and Istanbul Technical University Department of Control and Computer Engineering in 2001. He has been working as a research assistant in computer science department in Informatics Institute of ITU since his graduation.