Modeling studies for the detection of bacteria in Biosensor Water Distribution Networks

ANTONIO BERTOLDI

Master's Degree Project
Stockholm, Sweden  September 2012

XR-EE-RT 2012:031
Contents

1 Introduction 3

2 Background 5
   2.1 Bacteria and Biofilms ............................... 5
   2.2 Characterization of Biosensors ..................... 7

3 Wireless Sensor Detection systems 13
   3.1 Protocol Architecture ........................... 13
   3.2 Constraints imposed by protocol .................. 19
   3.3 Distributed detection over biosensors arrays .... 20

4 Investigation of Wireless Sensor Networks in water distribution systems 29
   4.1 Sensors for detection of E.coli .................... 29
   4.2 Probability of detection inside water network distribution systems 31
   4.3 Optimal Sensor Placement Problem ............... 32

5 Proposed Optimal Sensor Placement Algorithm 35
   5.1 Optimization Problem under economic constraints .... 35
   5.2 Network connection test ............................ 37
   5.3 Proposed Algorithm ............................... 40
   5.4 Computational complexity and speed up solution .... 42

6 Testing the solution 44
   6.1 Network topologies ............................... 44
   6.2 Simulations details and optimal sensor placement results ... 47
   6.3 Contamination scenarios ........................... 53

7 Conclusions 63
Abstract

The detection of bacteria in the water is a slow process that requires the use of expensive equipment and qualified personnel. However, real time fast detection is essential in water distribution networks. In this thesis we study the deployment of a wireless network of biosensors in a water distribution system, in order to detect contamination of a particular kind of harmful bacteria, the E.coli. This network will efficiently utilize the interconnected biosensors and achieve real time and in-field detection of the bacteria. Because of the non existence of biosensors hardware equipped with radio receivers and transmitters, we study theoretically the modeling of such a system and its potential application in real water distribution networks. The main goal of our study is to find an optimal sensor placement strategy to maximize the probability of detection, having a fixed number of sensors that must be placed in a connected topology. We propose a simple algorithm that solves the optimal sensor placement problem. The performance of the proposed approach have been evaluated by considering three different topologies simulated by the system simulator EPA NET. The simulation results show that the proposed algorithm provides the higher detection probability in the network compared to other solutions, such as random sensor placement.
Chapter 1

Introduction

The world of bacteria is made up of a multitude of different species, some of them useful and some harmful (pathogenic bacteria) for human. They can live in several ambient such as water or food, and growth in different temperature. When present in water, harmful bacteria cause several and different problems to human health, from diarrhea up to death. For this reason, a continuous water monitoring is required to avoid infection and damage caused by pathogenic bacteria. Nowadays their detection in drinking water is performed by the use of analytical methods, even made in laboratory and that require hours before return a result. These methods may involve also the use of expensive equipment and the presence of skilled personnel. Moreover in all these methods samples of water are collected and taken to the laboratory, preventing a complete monitoring of the water distribution system. The simpler way to have a detection system that works real-time and in-field is realize a wireless sensor network located within the water distribution system and which could control the water quality. This is suggested by the fact that these kinds of networks are used for monitoring some water characteristics like water leakages, temperature, pH and viscosity.

A fast, real-time and automated detection system, may reduce the time limitation of the classical and analytical methods, work directly in-field without the necessity to bring samples to laboratories and react faster when contamination is detected. At the base of these networks there is the sensor unit, an embedded system able to monitor and detect some environmental aspect, manage the local detection data and then communicate with other sensors.

It is possible to find on the market for different types of sensor for several kind of detection, specific set of them are the biosensors. A biosensor is an analytical device for the detection of an analyte that combines a biological component with a physicochemical detector. Generally is made up of three parts, a sensitive biological element that can react with the analyte, a transducer part that transforms the signal resulting from the interaction of the analyte with the biological element into another signal that can be more easily measured and quantified, and the biosensor reader device with the associated electronics or signal processors that are primarily responsible for the display of the results in a user-friendly way. Several different types of biosensors is actually used for the detection of bacteria in water or in food, so the idea of realize a wireless sensor network using this particular hardware is an open challenge that can
bring many benefits and reduce the complexity of the detection of bacteria in water distribution systems.

Scope of this thesis is to investigate the real possibility to have a wireless sensor network inside a water distribution system for monitoring the contamination of a particular kind of bacteria, the *E. coli*. The pathogenic form of *E. coli* cause serious food poisoning in humans, and are occasionally responsible for product recalls due to food contamination. The reservoir of this pathogen is mainly cattle or ruminants such as sheep and goats. Fecal contamination of water will lead the proliferation of these bacteria into the water distribution system. The strong correlation between *E. coli* and water fecal contamination is proved by that the presence of these bacteria in water is a test used to estimate this kind of water pollution.

In the thesis work the technology that best allows the detection of these bacteria in water pipes has been addressed, and then the availability of commercial biosensors of this type equipped with hardware able to create a biosensor network has been investigated. Since no commercially available hardware is found for this kind of detection, in the rest of thesis a complete description of how to realize this network is studied. At first which elements of a water distribution system may be the best for inserting sensors in them, and then which are the environmental and technical elements that can damage detection within nodes is described. The last aspect suggests to provide a way to reduce leakages of performances, so it is proposed an optimization problem in which the intent is to maximize the probability of detection of the entire system, giving as constraints a limited number of sensors, and the limitation of the distance between sensing nodes, in order to provide a wireless communication. The solution proposed is tested with three networks and are shown the benefit of using the provided architecture as a layout of a wireless sensor detection system for water monitoring.
Chapter 2

Background

In this chapter we introduce some elements of background useful for the thesis. The first part consists on an overview on the bacteria world. It explained how it is grouped and classified and why it is important to detect them inside a water distribution system. The second part consists on an introduction of the biosensors and it is described of which parts and function they have and how they work.

2.1 Bacteria and Biofilms

Bacteria are a large domain of prokaryotic microorganisms of size usually of the order of micrometers. The prokaryotes are a group of organisms that lack a cell nucleus (= karyon), or any other membrane-bound organelles nucleus *(prokaryote* comes to Greek προ- (pro-) “before” + καρυόν (karyon) “kernel”). In this kind of organism, neither their DNA nor any of their other sites of metabolic activity are collected together in a discrete membrane-enclosed area. Instead, everything is openly accessible within the cell. Bacteria have a wide range of shapes. Bacteria are present in most habitats on Earth, growing in soil, radioactive waste, water, and deep in the Earth's crust, as well as in organic matter and the live bodies of plants and animals, providing outstanding examples of mutualism in the digestive tracts of humans, termites and cockroaches. There are typically 40 million bacterial cells in a gram of soil and a million bacterial cells in a milliliter of fresh water; there are approximately five nonillion (5 × 10^{30}) bacteria on Earth, forming a biomass that exceeds that of all plants and animals. Bacteria are vital in recycling nutrients, with many steps in nutrient cycles depending on these organisms, such as the fixation of nitrogen from the atmosphere and putrefaction. In the biological communities surrounding hydrothermal vents and cold seeps, bacteria provide the nutrients needed to sustain life by converting dissolved compounds such as hydrogen sulphide and methane. Most bacteria have not been characterized, and only about half of the phyla of bacteria have species that cannot be grown in laboratory conditions.

Bacteria have a wide range of shapes, according with this property we can classify bacteria in:

- *Bacilli* (rod shaped)
• **Cocci** (spherical shaped)
• **Spirilla** (spiral shaped)
• **Spirochetes** (helical shaped)
• **Vibrios** (curved rod shaped)

Another way to subdivide bacteria is grouping them by the environmental temperature in which they live and grow. In this case we can distinguish three kinds of bacteria:

• **Cryophiles** or **Psychrophiles**: are capable of growth and reproduction in cold temperatures, ranging from $-15^\circ C$ and $10^\circ C$.
• **Mesophiles**: grow in moderate temperatures, typically between $20^\circ C$ and $45^\circ C$.
• **Thermophiles**: thrives at relatively high temperatures, between $45^\circ C$ and $122^\circ C$.

There are several methods used for bacteria identification, the most important is the Gram Staining method. Gram staining (or Gram’s Method) is a method of differentiating bacterial species into two large groups (**Gram-positive** and **Gram-negative**). It is based on the chemical and physical properties of their cell walls. Primarily, it detects peptidoglycan, which is present in a thick layer in Gram positive bacteria. A Gram positive results in a purple/blue color while a Gram negative results in a pink/red color. The main difference between Gram-positive and Gram-negative bacteria is that a Gram-positive bacterium has a higher amount of peptidoglycan in the cell wall, than Gram negative bacterium. In a Gram-negative bacterium, peptidoglycans constitute the 95% of the entire cell wall, in a Gram-negative only the 10%. A Gram-positive bacterium has only an inner membrane; instead a Gram-negative bacterium has two membranes the inner membrane, composed by peptidoglycans, and the outer membrane, composed by phospholipids and lipopolysaccharides.

An important kind of bacteria for human health is the Gram-negative, rod-shaped bacterium Escherichia coli. This bacterium is commonly found in the lower intestine of warm-blooded organisms (endotherms). Most E.coli strains are harmless, but some stereotypes can cause serious food poisoning in humans, and are occasionally responsible for product recalls due to food contamination. The harmless strains are part of the normal flora of the gut, and can benefit their hosts by producing vitamin $K_2$, and by preventing the establishment of pathogenic bacteria within the intestine. E. coli and related bacteria constitute about 0.1% of gut flora, and fecal-oral transmission is the major route through which pathogenic strains of the bacterium cause disease. Cells are able to survive outside the body for a limited amount of time, which makes them ideal indicator organisms to test environmental samples for fecal contamination. Optimal growth of E. coli occurs at $37^\circ C$, but some laboratory strains can multiply at temperatures of up to $49^\circ C$. Escherichia coli encompasses a enormous population of bacteria that exhibits a very high degree of both genetic and phenotypic diversity. Genome sequencing of a large number of isolates of E. coli and related bacteria shows that a taxonomic reclassification would be desirable. However, this has not been done, largely due to its medical importance and E.
coli remains one of the most diverse bacterial species: only 20% of the genome is common to all strains. Some of this could be very dangerous for human health, for example E. coli O157: H7 could cause diarrhea with abdominal cramps, followed by other severe organ system damage, including kidney failure. Moreover, E. coli bacteria easily spread from person to person, in particular when infected adults and children fail to adequately wash their hands. Similarly, restaurant workers not washing their hands when using the bathroom can pass on E. coli bacteria to food.

In water distribution systems it is possible to detect the presence of bacteria of E. coli inside of much more complex structures in which this type of pathogens coexist with other bacteria and form sort of colony. When contacting a solid surface, bacteria lay down gel-like polysaccharide matrix, which can trap other bacteria, forming a sort of colony, the biofilms. Thus, biofilms are structured groups of one or more microbial species encased in an extracellular polysaccharide matrix and attached to a solid surface. Some advantages of forming biofilm, in the perspective of bacteria are:

- increased availability of nutrients for growth
- increased binding of water molecules, which reduces the possibility of desiccation.
- protection against UV radiation, perhaps also physical protection. Biofilms protect microorganisms from antimicrobial agents.
- establishment of complex consortia, which allows for the recycling of substances.
- easier genetic exchange due to the proximity to progeny and other bacteria

In presence of fast water flows, biofilm clusters tend to become elongated in the flow direction to form filamentous streamers. This because with this kind of shape, the fluid forces which biofilm surface experiences is lower than other kind of shapes. At last, in this case the fluid flow determines an oscillatory movement of biofilms. It is possible with the high pressure of the water pipelines that biofilms are detached to the pipe surface (in which naturally lie without water flow) and are dragged by water through the entire pipelines. Developing a network for the detection of E.coli in water we have to take care of all the proprieties of bacteria and of biofilms described in this paragraph. In the next section an overview about the biosensors, units that let us detect bacteria in different situations and in to real-time constraints is provided. The existence of biosensor for the specific detection of E.coli is discussed in the next sections.

### 2.2 Characterization of Biosensors

A biosensor is a device for the detection of an analyte that combines a biological component with a physicochemical detector component. It is usually made up of 3 parts:

1. The *sensitive biological element or bio-marker*, some biological material (e.g. tissue, microorganisms, organelles, cell receptors, enzymes, antibodies, nucleic acids) that interacts (binds or recognizes) with the analyte
under study. The biologically sensitive elements can be created by biological engineering.

2. The transducer or the detector element, works in a physicochemical way; (optical, piezoelectric, electrochemical, etc.) and transforms the signal resulting from the interaction of the analyte with the biological element into another signal (i.e., transducers) that can be more easily measured and quantified.

3. The biosensor reader device with the associated electronics or signal processors, which is primarily responsible for the display of the results in a user-friendly way.

It is possible to group biosensors according to the sensitive element used or their transduction element. Biological elements used as sensitive part of biosensors are enzymes, antibodies, micro-organisms, biological tissue, and organelles. Antibody-based biosensors are also called immunosensors. Enzymes are proteins with high catalytic activity and selectivity towards substrates. They are very available in high purity levels in commerce, but their activity is strongly affected by several factors like pH, ionic strength, chemical inhibitors, and temperature. This kind of sensitive element is used coupled with electrochemical or fiber optic transducers. Enzymes have been immobilized at the surface of the transducer by adsorption, covalent attachment, entrapment in a gel or an electrochemically generated polymer, in bi-lipid membranes or in solution behind a selective membrane [3] [4]. Also antibodies are proteins, they are ideal for binding their antigen, for this immunosensors an outstanding selectivity. Also antibodies are largely commercially available, but for immobilizing these on a biosensor is required some step of treatment on it. They share similar limitations with enzymes, but they provide a faster and in-field detection of the analyte.

Other restrictions are that binding may not be reversible and the regeneration of the surface has very strong constraints (low pH, high ionic strength, etc). Antibodies are usually used with fiber optic or acoustic transducers, into low cost and single use sensors [2]. The use micro-organisms as biological elements in biosensors consist on the electrochemical measure of their metabolism, usually accompanied by the consumption of oxygen or carbon dioxide. Microbial cells are cheaper and more stable than enzymes or antibodies, anyway they are less selective and have long time period of recovery and response. Microorganisms have been immobilized, for example, in nylon nets, cellulose nitrate membranes, acetyl cellulose, or more recently into polycarbonate membranes[5]. The biosensor is described as an affinity sensor when the binding of the sensing element and the analyte is the detected event, is described as a metabolism sensor when the interaction between the biological element and the analyte is accompanied or followed by a chemical change in which the concentration of one of the substrates or products is measured. At last, when the signal is produced after binding the analyte without chemically changing it but by converting an auxiliary substrate, the biosensor is called a catalytic sensor [2].

Based on the sensitive element is possible to make another subdivision according to the kind of reception. If the sensitive element does not affect or change the target, the reception method is called bioaffinity-based reception, otherwise if the sensitive element catalyzes a bio-chemical reaction the reception is
called *bio catalytic reception*. Antibodies are bioaffinity receptors, enzymes both bioaffinity or biocatalytic receptors.

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzymes</td>
<td>Bioaffinity/Biocatalysis</td>
</tr>
<tr>
<td>Antibody</td>
<td>Bioaffinity/Immunosensor</td>
</tr>
<tr>
<td>Nucleic Acids/DNA</td>
<td>Biocatalysis</td>
</tr>
<tr>
<td>Biomimetic materials</td>
<td>Bioaffinity</td>
</tr>
<tr>
<td>Cellular Structures/Cells</td>
<td>Biocatalysis</td>
</tr>
<tr>
<td>Ionophore</td>
<td>Bioaffinity</td>
</tr>
</tbody>
</table>

Table 2.1: Types of biosensors based on receptors

Moreover there is two categories in which divide biosensors according to the kind of analyte detection

1. Label-free detection
2. Label-based detection

Label-based techniques require the labeling of query molecules with labels such as fluorescent dyes, radioisotopes, or epitope tags. This kind of detection may alterate the surface and natural activities of the query molecules, moreover labeling procedures are laborious and lengthy and they limit the number and types of analytes that can be monitored. Label-free detection methods depend on the measurement of an inherent property of the analyte (e.g. fluorescence or dielectric properties), these methods avoid interference due to tagging molecule, providing a faster way to detection.

Several transducer parts are used for have a biosensor unity, the most relevant are

1. Mechanical
2. Optical
3. Electrical
4. Piezoelectric
5. Electrochemical
6. Thermal

Examples of mechanical transduction methods are stress and mass sensing methods, are used in electro-mechanical devices such as the microcantilever biosensors. Past research works have reported the observation that when specific biomolecular interactions occur on one surface of a microcantilever beam, the cantilever bends [7][8]. In micro-cantilever biosensors there both stress sensing and mass sensing are used as sensing mode of detection. Stress sensing is
Figure 2.1: Detection principle of SPR device. Biomolecular interactions at the sensing surface layer are monitored as a shift in the resonance wavelength.

carried out by coating one side of the cantilever beam using a bio-receptor that adsorbs target biomolecules.

The adsorption results in the expansion or compression of the bio-receptor layer, which then induces surface stress on the cantilever beam, and thus the cantilever bends due to stress. In the latter type of sensing, the cantilever is actuated to vibrate in its resonant frequency. The binding of the biomolecule with a bio-receptor changes the frequency of vibration. The shift in resonant frequency is analyzed to detect the concentration of the biomolecule. Surface stress-based micro-cantilevers have been proposed and utilized because of their ease of operation, higher sensitivity, and the ease to study surface stress during adsorption through optical detection (as in atomic force microscopy (AFM)) and piezoresistive detection.

Due the velocity and the possibility to have in-field detection systems, optical transduction methods are the most used. Most diffused optical methods are the Fluorescence method, the Bioluminescence method and the Surface Plasmon Resonance (SPR). SPR is a surface-sensitive spectroscopic method that measures change in the refractive index of biosensing material at the interface between metal surfaces, usually a thin gold film (50-100 nm) coated on a glass slide, and a dielectric medium. When the surface plasmon wave interacts with a local particle or irregularity, such as a rough surface, part of the energy can be re-emitted as light that is possible to measure. In order to detect the analyte, the gold surface in SPR is immobilized with the sensitive element. When a binding between analyte and receptor happens, it is possible to detect it by measuring the angle of reflection of light of the SPR surface which is directly related to the amount of biomolecules bound to the gold surface. Figure 3.3 shows how a SPR sensor works.

Principal advantages of using SPR transduction are the real-time detection, and the possibility to have different biomarkers in order to have sensors capable to detect different analytes. Fluorescence and bioluminescence are based on the use of particular sensitive bio-receptors with bioluminescence or fluorescence properties. When the binding with analyte’s cells, the bioluminescence (or fluorescence) of these bio-receptors change, measuring this change is it possible to establish the amount of the analyte. The bacterium Vibrio Fischeri is an
example of bioluminescent sensitive element [6].

<table>
<thead>
<tr>
<th>Transduction Mechanism</th>
<th>Method</th>
</tr>
</thead>
</table>
| Mechanical             | Stress sensing  
                         | Mass sensing  |
| Optical                | Fluorescence  
                         | Chemiluminescence  
                         | Bioluminescence  
                         | Surface Plasmon  
                         | Evanescent Waves Interferometry  |
| Electrical             | Conductometric  
                         | Capacitive  |
| Piezoelectric          | Quartz Crystal Microbalance (QCM)  
                         | Surface Acoustic Wave (SAW)  |
| Electrochemical        | Potentiometric  
                         | Amperometric  
                         | Ion sensitive FET1 (ISFET)  
                         | Chemical FET (ChemFET)  
                         | Calorimetric  |
| Thermal                | Bioaffinity  |

Table 2.2: Biosensor classification based on transduction mechanism

Examples of electrical transduction methods are Conductometric or Capacitive methods, of electrochemical are the Potentiometric method and the Amperometric method. Amperometry is based on the measurement of the current resulting from the electrochemical oxidation or reduction of an electroactive species. It is usually performed by maintaining a constant potential at a working electrode (usually gold or carbon) or on an array of electrodes with respect to a reference electrode. The resulting current is directly correlated to the bulk concentration of the electroactive species. Potentiometric methods are based on measurement of the potential differences between an indicator and reference electrode when there is no significant current flowing between them. At last the most used piezoelectric transduction methods are the Quartz Crystal Microbalance (QCM) and the Surface Acoustic Wave (SAW). QCM sensor works with mass load effect of crystal. When a substance is absorbed on the electrode of the crystal, frequency goes down equivalent for the mass amount. This is called as mass load effect and the relationship between shifted frequency and mass is defined by Sauerbrey equation

\[
\Delta f = \frac{-2f_0^2}{A\sqrt{\rho q\mu q}}\Delta m
\]

Using the equation, it is possible to measure the mass change, so the amount of analyte. A general summary of the transduction methods is offered by table 2.2

Actually several types of hardware is developed for providing interaction between end users and biosensors. Generally biosensor producers, offer complex
systems in which the sensing part is already coupled with the hardware and the software for data management. For building smart detection systems with biosensors is important to provide them the equipment that make them able to communicate each other, and permit data transmission inside complex systems. It is also important to provide to the systems rules and protocols that make the communication easier and rigorous. For all these aspects in this work we suggest wireless technology as basic step for building this complex system. In the next chapter we describe the most important aspects of this technology, and in which way is possible to build smart detection system with it.
Chapter 3

Wireless Sensor Detection systems

In this chapter we explain how to build a distributed detection system using the wireless sensor technology. We first describe the protocol architecture of a Wireless Sensor Network. Then we introduce some constraints imposed by the protocols and by the nature of sensors. At the end explained the rules that are at the basis of distributed detection systems. For a general overview on estimation and detection over Wireless Sensor Network, see [18] and [19]

3.1 Protocol Architecture

A wireless sensor network (WSN) is a network with a distributed architecture, realized by a set of autonomous electronic devices able to monitor the surrounding environment and communicate each other. It is seen as a set of nodes with cheap hardware (no high value of RAM and CPU with low performances), called sensors or motes. These nodes, dispersed into the region of interest, are able to monitor some environmental effect and periodically send collected data to a central point of the network, called base station or gateway, that manages the network, collects data and can forward them to a remote system. Based to the hardware provided at each node, it is possible to implement controls and applications, for example for manage actuators or control systems, directly inside the nodes of the WSN. The basic components of a network for a system of this type are:

1. A set of distributed sensors
2. The central data fusion point
3. A set of software and hardware that permit the interaction between human and WSN.

The standard which specifies the physical layer and media access control for these networks is the IEEE 802.15.4 [20]. It is the basis for the standards that extend it by developing the not defined upper layers. As modulation technique for data transmission is used the Direct-Sequence Spread Spectrum (DSSS) technique. With DSSS, data being transmitted are multiplied by a "noise" signal,
a pseudorandom sequence of 1 and -1 values, at a frequency much higher than that of the original signal. The result is a noise-like signal, that can be used to reconstruct the original data at the receiving end, by multiplying it by the same pseudorandom sequence. This process is also known as de-spreading and mathematically constitutes a correlation of the transmitted pseudorandom number sequence with the pseudorandom number sequence that the receiver believes the transmitter is using. Timing between source and destination is required. A first advantage to use this technique is an enhancement of the Signal-Noise Ratio (SNR) on the channel, called also process gain. If an undesired transmitter transmits on the same channel but with a different pseudorandom number sequence (or no sequence at all), the de-spreading process results in no processing gain for that signal. This effect is the basis for the Code Division Multiple Access (CDMA) property of DSSS, which allows multiple transmitters to share the same channel within the limits of the cross-correlation properties of their pseudorandom number sequences.

There are three possible unlicensed frequency bands that are allowed for transmission:

1. 868.0-868.6 MHz in Europe, one communication channel allowed
2. 902-928 MHz in North America and up to ten channels allowed with channel spacing
3. 2400-2483.5 MHz, in worldwide use, up to sixteen channels with channel spacing

More than one network can coexists in the same area by using Frequency Division Multiplexing (FDM). Other functions defined at the PHY layer of this standard are:

1. Transmission and reception of a bit at physical layer
2. Turn on/off radio transmission equipment (required for the energy preserving problem).
3. Energy detection: measure the signal power, used for channel selection.
4. Link Quality Indication (LDI): characterization of the received signal based on his power and quality.
5. Channel selection.
6. Clear Channel Assessment: used for verify if a channel is free or is already used for a transmission.

Figure 3.1 shows the Physical Protocol Data Unit defined for the the IEEE 802.15.4 standard. It consists mainly in three parts. The Synchronization Header (SHR) permit timing between transmitter and receiver; the Physical Header (PHR) contains information about the frame such as the length, and the Payload that contains the MPDU (Mac Protocol Data Unit), this part may have a variable length.
Two types of network nodes are defined by standard, the *Full-Function Device* and the *Reduced-Function Device*. A Full-Function Device can be used as coordinator of the personal area network just as it may works as a common network node. It implements the general model for communicating with other nodes, and at the same time it is able to rely data of other nodes, in that case this node take the function of coordinator, when a node is a coordinator for the entire network it is also named *PAN (Personal Area Network) coordinator*. Reduced-Function Device is a very simple kind of node, with modest resources and communication requirements that can only communicate with FFD nodes. According to this, different topologies is provided, the most simple are:

1. **Star topology networks**

2. **Peer-to-peer (Mesh) topology networks**

Figure 3.2 shows an example of these topologies. In the first there is a central node, that have a rule of PAN coordinator and all the other nodes can communicate only with it. The PAN coordinator is always a FFD device; conversely the other nodes of the network should be either FFD or RFD devices. Peer to peer topology is a more complex network in which the communication with nodes is no forced like the star topology. Peer-to-peer (or point-to-point) networks can form arbitrary patterns of connections, and their extension is only limited by the distance between each pair of nodes. Since the standard does not define a network layer, routing is not directly supported, but such an additional layer can add support for multi-hop communications. A more complex topology is the cluster tree topology. In this topology, sensors are grouped into clusters; all of them have a coordinator and may be either a star or a peer to peer network. Each cluster have a node the cluster head that can communicate with the PAN coordinator. It is possible to see this as a tree with the PAN coordinator as root, only the cluster heads as root’s child and then the other nodes of the network.
Two operating modes are possible for the 802.15.4 MAC layer

1. Beacon Enabled
2. Beaconless

Beacon is a special control frame sent by the coordinator to their clients in order to synchronize them. All clients are listening and waiting for beacon frames, if they don’t receive this frame for long interval of time they pass to sleep-mode, in order to avoid battery consumption. Beacon frames are important in topologies like the star or the cluster tree for keep all nodes synchronized each other. The MAC method used is the CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance). In this method nodes transmit data only when the channel is sensed to be idle. With Collision Avoidance after start to transmit data a node send a packet for request the possibility to send them, the (*Request
to Send frame), only when the destination send back a frame of authorization to send (Clear to Send frame), it start the transmission of the data. There are 4 different types of frame:

1. Data Frame
2. ACK Frame
3. MAC Control Frame
4. Beacon Frame

The data frame have a variable length, the maximum number of byte of this frame is 128. The ACK frame is a short packet and it is sent as confirmation of the correct message reception. The control frame is used for control and configuration of clients. As said above beacon frames are sent by coordinator in order to synchronize all the clients. Additionally, a superframe structure, defined by the coordinator, may be used, in this case two beacons act as its limits and provide synchronization to other devices as well as configuration information. A superframe consists of sixteen equal-length slots, which can be further divided into an active part and an inactive part, during which the coordinator may enter power saving mode, not needing to control its network. Beaconless Mode there is no beacon frames periodically transmitted. The PAN coordinator sends a beacon frame only when new clients request to join the network.

IEEE 802.15.4 specifies only PHY and MAC layers of Wireless Sensors Networks. Higher levels are developed by other standards; the most used of them is ZigBee. This standard is developed for reduce power consumption, and the economic costs of the nodes, these characteristics make it the best fitting standard for Wireless Sensor Networks. It provides routines from Network to Application of the OSI protocol suite. At level Network the skills provided by ZigBee are:

1. Complete the 802.15.4 MAC layer with the higher parts of the MAC level.
2. Selection of network rules like Coordinator, Router, End-Device.
4. Implementation of Network layer address, packet and control.
5. Management of security keys.
6. Routing

Network addresses provided by ZigBee are of 16 bit of length, and permit sensors to communicate at level Network of the OSI model. ZigBee routing may be in two modes:

1. Tree Routing
2. Table Routing

In the first type, packets are forwarded to the child, or the father, of the destination node according with 802.15.4 primitives. In Table Routing a preliminary cycle of discovery permits to build a routing table by the transmission of a broadcast packet. Whatever is the topology of network, there is always
inside the nodes the Coordinator, a Full-Function Device which main functions are to initialize the network, to give the possibility to end devices to connect to this and provide to the management, the security of the entire system. Another important device inside a ZigBee Network is a Router. This kind of node is a Full-Function Device that allows to the End-Devices to communicate each other. The way of forwarding packets depends on the topology of the network. At last the simplest nodes available for a ZigBee network are the End-Devices, that are generally Reduced-Function Devices. These nodes should be in sleep or active mode, if it is in the first mode the Router that can communicate with it must save all massages having it as destination. ZigBee standard permits the constitution of three different topologies of networks: star, mesh and tree networks. The most simple is the star topology (figure 3.3). This network is made up of a FFD with the role of Coordinator and a set of End Devices that can communicate only with it. This network is not very scalable and when the number of end devices increases, increases also the number of data that the coordinator have to manage, this may affect the network performances.

Figure 3.3: Example of ZigBee star topologies

The second alternative is the cluster tree topology (see figure 3.4 (b)). In this topology the coordinator is placed on the root of the network, than there is some simple rules to build the rest of network:

1. Coordinator must be connected either to a Router or to End Devices.
2. Child of a Router should be either Router or End Devices.
3. End Devices must have not child

In this topology communication is allowed only between child and father, for sending a message it is necessary to go up into the tree until the first ancestor of the destination node is reached, than it is possible to go down through his child until the destination node. An important problem of this topology is that no redundancy is provided, so if a node does not work properly, the entire set of nodes below became unreachable. At last, the more complex Network provided by ZigBee standard is the Mesh network in which:

1. A Coordinator can communicate with his child, and with all nodes within his radio range.
2. A Router can communicate with his End-Device child and all Routers or the Coordinator if radio communication is possible.

This topology is more scalable, more secure and the possibility of each Full-Function Device to communicate with all the FFD inside his radio range provide to the topology a strong property of redundancy.

![Mesh Topology Network](image1)

![Cluster Tree Network](image2)

Figure 3.4: Examples of ZigBee Tree and Mesh topologies

### 3.2 Constrains imposed by protocol

The realization of a Wireless Sensor Network for distributed detection of bacteria must take into account some constraint derived from the nature of the protocol and the sensors hardware. Sensor nodes must be devices of small size due mobility constraints, they have to be autonomous and, especially in environmental monitoring, they usually are unattended and dispersed into the region of interest. In an ideal network any intervention of the human (e.g. maintenance or substitution of nodes) must be reduced, that means that sensors must have hardware very adaptive to the surrounded environment. A sensor node consumes power either for sensing, communicating or data processing. The most
amount of energy is spent for data communication, anyway ZigBee standard provide routines that may reduce the energy consumption for the communication process (e.g. "sleep" mode for RFF devices). Since the main source of power supply for sensor nodes are batteries, even on-rechargeable one, sensors must work taking into account the low-power consumption constraint. High portability, low dimensions, low energy consumption, high independence respect human maintenance, high adaptability may let became sensors expensive devices. On the other hand have an optimal monitoring, a certain number of sensors is required, since the amount of sensors requested may be a relevant number, sensors devices production cost must be limited.

Other constraints are given by communication skills. At first, Wireless sensors have a restricted radio range, given two nodes, in order to permit communication between them it is required that they are at distance lower than a communication threshold. Since the radio range for ZigBee standard is between 10 and 75 meters, the distances between two nodes that have to communicate each other must be in this range. Given the area that the Wireless Sensor Network have to control, the amount of sensors that it is required for having a satisfying monitoring has as lower bound the number of sensors dispersed in the area that is able to communicate. When the dimensions of the region of interest increase, also the number of required sensors is greater. Since the cost of sensors is always a constraint to take into account, it is important to find the optimal position of sensors in order to reduce costs. Other aspects that may damage communication are the presence of devices that create interference or physical obstacle that may affect signal transmission. Both the event described above may reduce the throughput and damage the network performances. In order to solve the problems described above, it is possible to place sensors using the mesh network (section 3.1) to provide redundancy of paths between FFD nodes. RFD nodes must be protected and inserted into the network in order to provide them an efficient connection with their coordinator.

In the next section we describe the theory of distributed detection systems, that represent the basis of the realization of wireless sensor networks for environmental monitoring.

### 3.3 Distributed detection over biosensors arrays

Given \( n \) sensors that monitor the same environmental aspect, related a related detection problem consist on give a global and unique answer to the state of the nature, based to their local reports. In the most simple definition of the problem, it consist on decide if the monitored event is happened or not, based on the signal received from transducers. Mathematically, let be \( H_0 \) and \( H_1 \) two hypothesis defined as

\[
\begin{align*}
H_0 &: \text{ Desired signal absent} \\
H_1 &: \text{ Desired signal present}
\end{align*}
\]

a distributed detection system establishes which of the two event is verified, on the basis of the signals received from sensors.

There are several architectures of a distributed detection system; the most used is the parallel architecture with fusion center (see figure 3.5(a)). In this
architecture each sensor sends his local decision message to the fusion center that by combining these messages returns the global answer of the system. The principal problem is that solution is not salable, and when the number of the nodes increases, the fusion center has to manage more information, this damages network performances. Another possibility is to have a parallel architecture without fusion center (see figure 3.5(b)). In this case sensors transmit each other the results of their observation and reach a global answer in multiple transmission steps, until consensus between sensors is reached. This solution is more salable than the first, but reaching a global answer can take more time than the first, moreover the probability of error is bigger than the architecture with fusion center.

\[ S_i \rightarrow S_{i+1} \rightarrow \cdots \rightarrow S_n \]

Figure 3.5

In the serial case the sensor $S_i$ sends the output of his local decision to the sensor $S_{i+1}$ until the last sensor of the array. In this way the decision of the $S_n$ sensor corresponds to the final global answer of the system. This architecture have poor performance in terms of time than the parallel solution, and an error at the $i\text{-th}$-sensor may affect the solution of all the next sensors of the network. Moreover is it possible to organize the network as a tree (3.6(b)). Detector nodes is grouped in clusters and sent their messages to the cluster head, which operate a partial local fusion and send the result to the fusion center. This solution is more salable than the parallel, and detection errors of nodes can
be resolved at cluster-head level. This multi-hop architecture best resolve the problem of scalability for large networks, and also fit best with the organization of a wireless sensor network described topology described into section 3.1.

![Diagram of distributed detection system](image)

(a) Example of serial topology of a distributed detection system

(b) Example of a tree topology for a distributed detection system without Fusion Center

Figure 3.6

An important aspect for a distributed detection system is the choice of the optimal fusion data rule. Each sensor receives a signal $y_i$ from the monitored element, and based on his observation can estimate at which hypothesis class belongs $y_i$. After the decision, the sensor transmits a message $u_i$ in which describes his local decision, the final global decision is obtained by an opportune fusion of all local sensors decisions. It is possible to define the local decision $u_i$ with $i = 1, 2, ..., n$ as

$$u_i = \begin{cases} 
-1 & \text{if } H_0 \text{ declared} \\
+1 & \text{for } H_1 \text{ declared}
\end{cases} \quad (3.1)$$

Then, a global data fusion rule is defined as:

$$u_0 = f(u_1, u_2, u_3, ..., u_n) \quad (3.2)$$

The function $3.2$ is the general definition of a fusion rule for a distributed detection system. Data fusion rules are often implemented as "$k$ out of $n$" logical functions. That means that if $k$ or more detectors decide hypothesis $H_1$, then the global decision is $H_1$. We can define the output of the detection system $u$ as
\[ u_0 = \begin{cases} +1 & \text{if } u_1 + u_2 + u_3 + \ldots + u_n > 2k - n \\ -1 & \text{otherwise} \end{cases} \] (3.3)

Common logical functions such as AND, OR, are special cases of the k out of n rule. Several papers propose the Likelihood Ratio Test (LRT) as an optimum solution for the problem. The LRT formulation for distributed detection problems is

\[ \Lambda = \frac{P(u|H_1)}{P(u|H_0)} \] (3.4)

with

\[ u = [u_1, u_2, u_3, \ldots u_n] \]

According with the Neyman-Pearson lemma, we can use LRT to obtain the optimum fusion rule. Let \( u_0 \) be the final global decision, then

\[ u_0 = \begin{cases} +1 & \text{if } \Lambda > T \\ -1 & \text{otherwise} \end{cases} \] (3.5)

Where T is a threshold value for the detection system. For \( \Lambda = T \), according with the Neyman-Pearson formulation we set \( u_0 = 1 \) with probability \( \epsilon \). The values of \( T \) and of \( \epsilon \) are chosen for having the probability of false alarm of the system \( P_f < \alpha \) and maximize the probability of detection \( P_d \).

If the sensors observation are conditional independent\[^{10}\], the LRT ratio may assume a different and more easy form. For the conditional independence property

\[ P(u|H_1) = \prod_{i=1}^{n} P(u_i|H_1) \]

and

\[ P(u|H_0) = \prod_{i=1}^{n} P(u_i|H_0) \]

So, the LRT test \( \Lambda \) under conditional independence became:

\[ \Lambda = \prod_{i=1}^{n} \frac{P(u_i|H_1)}{P(u_i|H_0)} \] (3.6)

In their paper, Z. Chair and P.K. Varshney\[^{9}\] use the Bayes optimum threshold as T, so the (3.5) becomes:

\[ u_0 = \begin{cases} +1 & \text{if } \Lambda > \frac{P_0(C_{01} - C_{00})}{P_1(C_{10} - C_{11})} \\ -1 & \text{otherwise} \end{cases} \] (3.7)

Where \( P_0 \) and \( P_1 \) are the a priori probabilities of the two hypotheses

\[ P(H_0) = P_0 \]
and $C_{ij}$ denotes the cost of global decision being $H_i$ when is present $H_j$. In the minimum probability of error criterion case, that is, $C_{00} = C_{11} = 0$, and $C_{10} = C_{01} = 1$ the (3.7) threshold becomes:

$$T = \frac{P_0(C_{01} - C_{00})}{P_1(C_{10} - C_{11})} = \frac{P_0}{P_1}$$

Using Bayes rule to express the conditional probabilities, multiplying $\Lambda$ with $T^{-1}$, Z. Chair and P.K. Varshney obtain:

$$\frac{P(u|H_1) P_1}{P(u|H_0) P_0} = \frac{P(H_1|u)}{P(H_0|u)}$$

(3.8)

With this result the fusion rule becomes:

$$u_0 = \begin{cases} +1 & \text{if } \frac{P(H_1|u)}{P(H_0|u)} > 1 \\ -1 & \text{otherwise} \end{cases}$$

(3.9)

The corresponding log-likelihood ratio test is:

$$u_0 = \begin{cases} +1 & \text{if } \log \frac{P(H_1|u)}{P(H_0|u)} > 0 \\ -1 & \text{otherwise} \end{cases}$$

(3.10)

Z. Chair and P.K. Varshney propose the sequent optimal fusion rule:

$$f(u_1, u_2, ..., u_n) = \begin{cases} +1 & \text{if } a_0 + \sum_{i=1}^{n} a_i u_i > 0 \\ -1 & \text{otherwise} \end{cases}$$

(3.11)

where $a_i$ are the optimal weight defined as:

$$a_0 = \frac{P_1}{P_0}$$

$$a_i = \begin{cases} \log \frac{1-P_{M_i}}{P_{F_i}} & \text{if } u_i = 1 \\ \log \frac{1-P_{F_i}}{P_{M_i}} & \text{otherwise} \end{cases}$$

where $P_{M_i}, P_{F_i}$ are the probabilities of missing and false alarm of each sensor. 

The limitation of the approach offered by Z. Chair and P.K. Varshney is that the a priori probabilities even are not known. In addition when observation do not satisfy the conditional independency, estimate the value of $\Lambda$ may be difficult. A possible enhancement of the Z. Chair and P.K. Varshney result is provided by Jian-Guo Chen and Nirwan Ansari [11]. In their work they describe an adaptive and incremental algorithm as decision fusion rule of a system where local decisions are conditionally dependent. At the basis of his algorithm there is a new simplification of the equation 3.4.
\[ \Lambda = \frac{P(u|H_1)}{P(u|H_0)} = \frac{P_1(u_1)P_1(u_2|u_1) + \ldots + P_1(u_k|u_1, u_2, \ldots, u_{k-1})}{P_0(u_1)P_0(u_2|u_1) + \ldots + P_0(u_k|u_1, u_2, \ldots, u_{k-1})} \]  

(3.12)

where \( P_i(u_k|u_1, u_2, \ldots, u_{k-1}) = P(u_k|u_1, u_2, \ldots, u_{k-1}, H_i), i = 0, 1, \) are the conditional probability of the local decision \( u_k \) given the hypotheses \( H_i \) and the local decisions of the other sensors. In their paper Jian-Guo Chen and Nirwan Ansari show that:

\[ P_i(u_k|u_1, u_2, \ldots, u_{k-1}) = \frac{1}{P_i(u_1, u_2, \ldots, u_k) + P_i(u_1, u_2, \ldots, u_k)} \]  

(3.13)

then, after posed

\[ p_k = \frac{P_1(u_1, u_2, \ldots, u_{k-1}, u_k = -1)}{P_1(u_1, u_2, \ldots, u_{k-1}, u_k = +1)} \]

and

\[ q_k = \frac{P_1(u_1, u_2, \ldots, u_{k-1}, u_k = -1)}{P_1(u_1, u_2, \ldots, u_{k-1}, u_k = +1)} \]

they propose as fusion rule the equation

\[ f(u) = \sum_{i=0}^{N} W_i u_i \]  

(3.14)

where \( W_i \) is the weight associated to the decision of the \( i \)th-sensor obtained as:

\[ W_0 = \log \frac{P(H_1)}{P(H_0)} \]

\[ W_1 = \begin{cases} 
\log \frac{P_1(u_1)}{P_0(u_1)} & \text{if } u_1 = 1 \\
\log \frac{P_0(u_1)}{P_1(u_1)} & \text{otherwise}
\end{cases} \]

and for \( k > 1 \):

\[ W_k = \begin{cases} 
\log \frac{1+q_k}{1+p_k} & \text{if } u_k = 1 \\
\log \frac{q_k(1+p_k)}{p_k(1+q_k)} & \text{otherwise}
\end{cases} \]

where \( p_k, q_k \) is defined as above. Conditional independence of observation is not require with the fusion rule proposed, but the a priori probabilities \( P_0 \) and \( P_1 \) still remain a problem.

In the last part of their work, Jian-Guo Chen and Nirwan Ansari show how avoid the lack of knowledge about these probabilities. Let be \( n \) the number that \( H_1 \) occurs, \( m \) the number that \( H_0 \) occurs, and \( m_{k,1} \) the number of \((u_1, u_2, \ldots, u_1, u_{k-1}, u_k = +1, H_1)\) occurs.
The number of \((u_1, u_2, \ldots, u_{k-1}, u_k = -1, H_1)\) occurs

\(n_{k,1}\) the number of \((u_1, u_2, \ldots, u_{k-1}, u_k = +1, H_0)\) occurs

\(n_{k,0}\) the number of \((u_1, u_2, \ldots, u_{k-1}, u_k = -1, H_0)\) occurs

Then, the values of \(p_k\) and \(q_k\) may be approximated as

\[ p_k \approx \frac{m_{k,0}}{m_{k,1}} \]
\[ q_k \approx \frac{n_{k,0}}{n_{k,1}} \]

So, based on this approximation the weights \(W_k\) became:

\[ W_k = \log \frac{1 + q_k}{1 + p_k} \approx \frac{n_{k,1} + n_{k,0}}{n_{k,1} m_{k,1} + m_{k,0}} \]

when \(u_k = +1\). Here it is possible to note that \(n_{k,1} + n_{k,0} = n_{k-1,j}\) and \(m_{k,1} + m_{k,0} = m_{k-1,j}\) where:

\[ j = \begin{cases} 1, & \text{if } u_{k-1} = +1 \\ 0, & \text{otherwise} \end{cases} \]

At this point it easy to understand that

\[ W_k \approx \log \frac{m_{k,1}}{n_{k,1}} - \log \frac{m_{k-1,j}}{n_{k-1,j}} \]

if \(u_k = +1\) and

\[ W_k \approx \log \frac{n_{k-1,j}}{m_{k-1,j}} - \log \frac{m_{k,0}}{n_{k,0}} \]

Jian-Guo Chen and Nirwan Ansari presented a solution on how use the approximation for avoid the not knowledge of \(P_0\) and \(P_1\). Based on their approach there is the concept of reinforcement learning\(^{[12]}\), that consists on that if the current local decision of the sensor \(k\) conforms to that of the fusion center, his weight is reinforced, otherwise should be reduced. The amount of reinforcement and reduction is given by the partial derivation of \(W_k\) with respect to \(m_{k,0}, m_{k,1}, n_{k,0}, n_{k,1}\). The reinforcement value is:

\[ \Delta W_k = \begin{cases} \frac{\partial W_k}{\partial m_{k,1}} = \frac{1}{m_{k,1}}, & \text{if } u_{k-1} = +1 \text{ and } H_1 \\ \frac{\partial W_k}{\partial n_{k,0}} = \frac{1}{m_{k,0} n_{k-1,j}} e^{-W_k}, & \text{if } u_{k-1} = -1 \text{ and } H_0 \end{cases} \]

In the same way is it possible to find that the reduction amount is:

\[ \Delta W_k = \begin{cases} \frac{\partial W_k}{\partial m_{k,0}} = \frac{1}{m_{k,0}}, & \text{if } u_{k-1} = +1 \text{ and } H_1 \\ \frac{\partial W_k}{\partial n_{k,1}} = \frac{1}{m_{k,0} n_{k-1,j}} e^{W_k}, & \text{if } u_{k-1} = -1 \text{ and } H_0 \end{cases} \]
An other fusion rule is provided by Zhi Quan and Shuguang Cui, that describe a linear fusion rule in which weights are obtained using an optimization problem [13]. At the basis of this proposal there is the assumption that the summary statistics from local sensors are normally distributed. Let be the vector \( \mathbf{u} \) at the fusion center a realization generated from an N-dimensional normal (Gaussian) distribution under each hypothesis.

\[
\mathbf{u} \sim \begin{cases} 
N(\mu_0, \Sigma_0), & H_0 \\
N(\mu_1, \Sigma_1), & H_1
\end{cases}
\]

where \( \mu_0(\mu_1) \) and \( \Sigma_0(\Sigma_1) \) are the mean vector and covariance matrix of \( \mathbf{u} \) under \( H_0(H_1) \).

Zhi Quan and Shuguang Cui try to maximize \( P_d \) with an upper limit on \( P_f \).

The proposed linear fusion rule is:

\[
T(\mathbf{u}) = \sum_{i=1}^{N} w_i u_i = \mathbf{w}^T \mathbf{u} \geq \gamma
\]

where \( \mathbf{w}^T = [w_1, w_2, \ldots, w_N]^T \) are the weight coefficients obtained with the optimization problem. The probabilities of false alarm are expressed as:

\[
P_f = P(T(\mathbf{u}) \geq \gamma | H_0) = Q(\frac{\gamma - \mathbf{w}^T \mu_0}{\sqrt{\mathbf{w}^T \Sigma_0 \mathbf{w}}})
\]

and

\[
P_d = P(T(\mathbf{u}) \geq \gamma | H_1) = Q(\frac{\gamma - \mathbf{w}^T \mu_1}{\sqrt{\mathbf{w}^T \Sigma_1 \mathbf{w}}})
\]

Where \( Q \) is the complementary cumulative distribution function. After express \( \gamma \) as:

\[
\gamma = \mathbf{w}^T \mu_0 + Q^{-1}(\varepsilon) \sqrt{\mathbf{w}^T \Sigma_0 \mathbf{w}}
\]

The proposed optimization problem proposed in [13] for obtain the \( \mathbf{w} \) is:

\[
\max_{\mathbf{w}} P_d = \max_{\mathbf{w}} Q(\frac{Q^{-1}(\varepsilon) \sqrt{\mathbf{w}^T \Sigma_0 \mathbf{w}} - (\mu_1 - \mu_0)^T \mathbf{w}}{\sqrt{\mathbf{w}^T \Sigma_1 \mathbf{w}}})
\]

For solving the optimization problem Zhi Quan and Shuguang Cui propose a semi-definite programming strategy. In this section it has been described how implement a distributed detection system. Several aspects emerged and need to be defined in the practical realization of a distributed detection system:

1. The topology of the system must be defined, according to the technology used for his realization
2. It is important to define if local decision are or not conditionally independent, in order to set-up the proper fusion rule
3. Threshold related to the global decision, and the upper bound of probability of false alarm must be defined based on the practical requirements of the aimed environmental monitoring system.
All these aspects were developed and described in the next chapter for the monitoring case of study, a water distribution system.
Chapter 4

Investigation of Wireless Sensor Networks in water distribution systems

In this chapter we discuss about the main characteristics of a wireless sensor network for the detection of bacteria inside a water distribution system. In the first part we describe the characteristics and the constraints of the biosensors that today is used for the detection of E.coli in water. Then we explain in which way place sensors inside nodes and the topology of a wireless sensor network for the detection of bacteria in water distribution systems. In the second part is described which environmental and practical element have an impact on the probability of detection inside the nodes of a wireless distribution system, at last where analyzed some related works about the optimal sensor placement in water distribution systems.

4.1 Sensors for detection of E.coli

The most used biosensors for the real time detection of E.coli in water is biosensors based on the Quartz Crystal Microbalance technology. As introduced in the section 2.2, QCM sensor works using the mass load effect of a crystal, that consist on the relationship between the resonance frequency of the crystal and the amount of the analyte mass present on the sensor surface and it is defined according the Sauerbrey equation as:

$$\Delta f = \frac{-2f_0^2}{A\sqrt{\rho q\mu_q}} \Delta m$$

Since in water coexists several kind of bacteria, in order to have a best detection E.coli antibody are used as capture element due the high sensitivity that they have with the analyte. Main limitations in the use of this kind of sensors in a detection system are:

1. The requirement of the direct contact between the E.coli bacteria and the biologic element used in the sensor.
2. The limitation on the water flow in which use this kind of sensors.

3. The necessity of a periodic maintenance of the sensors for remove the membrane inserted on it.

This sensors are today used in laboratories in which only some water samples are tested, the use of them in an *in-field* detection of bacteria may be difficult since the dimension of water tanks or pipes can make it difficult the binding between sensors and analyte. A possibility to reduce the impact of this limitation is to use several sensors inside a node of the water distribution system; in that case the binding of the bacteria and the array of sensors may be easier.

Another limitation of this sensors is that the binding between bacterium and antibody may require time after it is completed, so the water flow must have a speed slow to let this action completed. QCM sensors work in a flow velocity between 50 to 100 $\text{mu}l/min$. That means that the use of these sensors inside a water distribution system is impossible due water flow velocity. At last, these sensors require continuous maintenance due the need to change the used membrane of antibodies. This justifies why the analysis of the presence of these bacteria in the water today is performed only in specialized laboratories, and not directly into water systems. Sensors that detect this kind of bacteria with low constraints of the QCM are the optical sensors. For this kind of sensors it is inserted inside the water pipeline a membrane of some biological element with fluorescent properties (e.g., *Vibrio Fischeri*) that in proximity of the analyte are subject to a modification of this property, by measuring the difference in fluorescence is possible to determine the amount of bacteria. Advantages in using these sensors are the faster detection and the possibility to use it in *in-field* detection of bacteria. No commercial products of these sensors are available, optical sensors for detection of E.coli is only argument of academicals works. So a practical realization of a wireless sensor network using these sensors may be impracticable. The commercial availability of products it is a limitation also for the QCM sensors. It is possible to find several companies that provide them with or without the biological membrane required for the detection. The problem for these sensors is that they are provided or without any hardware that permits the transmission and the management of data or in complex systems in which are provided also a specific mode of transmission and management of the data, even with specific software for the analysis of the results of the detection.

First we investigate about how to place sensors inside a water distribution system, since the flow velocity and the size of a node may affect the probability of detection of sensors inside this system a correct placement of sensors may be the nodes of the system, due the limited flow velocity inside them. A suggestion to how to insert them is provided by the PIPENET project [14]. Scope of this work is to build a wireless sensor network inside water distribution system in order to detect leakages of water. A set of different sensors is placed inside a node, and a gateway is placed outside it in order to collect information and transmit them to others gateways of the network without leakages of communication performances. This type of positioning is recommended by the fact that in case there were no external gateway units, the communication between the nodes of the network could be damaged by the presence of means such as cement, plastics or the water itself, which could damage the performance of network communication. Using external gateway the leakages in communication
performances is limited due the proximity between the gateway and sensors, and also the best quality of communication obtained when there are no physical barriers that hinder communication. So to deploy a wireless sensor network for detection inside water distribution system the use of a topology in which sensors are placed inside a nodes and communicate with an external unit that collect information and transmit it through the rest of network is strongly recommended. The topology that is better suited for the realization of a network of this type is the units mesh (or cluster-tree) one, in which gateway are FFD units that collect data internal units, that can be both RFD or FFD, since the main function of the latter unit is to transmit data to the gateway, to reduce the costs of these units may be all RFD units, some of them may be FFD units if it is required to have redundancy of paths also in the cluster. At last, in order to implement a correct fusion rule, observations inside the same cluster are conditionally dependents, while observations of different clusters may be either dependent or independent, according to whether or not they share some path within the network. If two clusters share some water path inside the network, than their observations are conditionally dependent, in the other case are conditionally independent. In the next section we explain which environmental and practical characteristics of a node may affect the probability of detection inside it.

4.2 Probability of detection inside water network distribution systems

Many factors may influence the detection of E. coli in water distribution systems. Some of them are environmental factors (water exposure to animal feces and water temperature); others are specific features of the water distribution system (flow velocity, nodes or pipes volume).

Environmental factors may influence the emergence and proliferation of E. coli bacteria in water. The strong correlation between E.coli and water fecal contamination is proved by the fact that verify the presence of these bacteria in water is a test used to estimate this kind of water pollution. Another important environmental feature that promotes proliferation of E.coli in water is the temperature because E.Coli is a mesophilic bacterium that can grow in temperatures ranging from 7°C to 50°C, with an optimum of 37°C. For this, first information that can help to understand where to place sensors is verify the presence of livestock farms close a node of the water network and collect information regarding previous origins of fecal contamination into the system, or regarding the temperature that water can reach inside nodes.

Moreover the material and the size of node where array of sensors is placed may influence the probability of detection. Biofuels into which is possible to find E.coli strains have a natural hydrophobicity that increase with the dimension them. Therefore, the biofuel strains are forced towards the surface of a node and attack his surface settling on it. Due to the fact that the detection of these bacteria require the contact between the sensor binding part of and the strain, the selection of nodes with small volumes improve the probability of detection because in this is more likely to have this contact.

In order to have the best performances of detection, biosensors have an op-
timal water flow velocity threshold. The probability of miss detection increases in nodes with water flow velocity greater than the threshold value, moreover very fast flow velocity damages the biosensor hardware. The particular threshold varies from sensor to sensor, and depends mainly on the sensitive part used into them. For instance, Quartz Crystal Microbalance sensors have a velocity threshold lower than optical ones, because the QCM surface requires more time for binding the bacteria than the similar part of an optical device. Moreover, it is possible that a node is subject to a flow rate so high as to prevent the bacteria to move toward the border, dragging them across the node too quickly to permit it to attack to the node surface.

Thus, the probability of detection \( p_{di} \) associated to the node \( v_i \) depends to all this factors and without experiments is it difficult to give it a quantitative function.

Let is:

1. \( \sigma_i \) the value of the volume of the node
2. \( \omega_i \) the average of flow velocity into a node \( v_i \)
3. \( t_i \) the average of temperature into a node \( v_i \)
4. \( m_i \) the size of the node \( v_i \)
5. \( \gamma_i \) the material of which node \( v_i \) is made

A possible definition of probability of detection \( p_{di} \) is:

\[
p_{di} = f(\sigma, \omega, t, m, \gamma)
\]  

(4.2.1)

### 4.3 Optimal Sensor Placement Problem

We study the problem of finding the optimal way to put sensors into the water distribution system, in order to maximize the probability of detection of the system. Related works show that this problem can be formulated in different ways, since many possible objectives can be formulated for optimal sensor placement. The main objective of this works are minimize the density of population exposed to a contamination event, or maximize the quantity of monitored water into a water distribution network, the minimization of the risk associated with a contamination event. In order to find the optimal architecture of the sensor network, it’s required to have some preliminary information about the water distribution system. The first information is the topology of the network. A water distribution system is an oriented graph \( G = (V, E) \), where \( V \) is a set of vertices, or nodes, where pipes meet, and \( E \) is the set of edges representing pipes, the direction of the edges represent the way in which water flows in the network. Vertices can represent sources, such as reservoirs or tanks, where water is introduced, and sinks (demand points) where water is consumed. Each pipe connects two vertices \( v_i \) and \( v_j \) and is usually denoted \( (v_i, v_j) \). At each edge is associated a weight, which identify the volume of the water that flow through the edge per unit of time. At second, it is required the amount of water demand at each node. At last different information is used by related works for the proposal of their optimization problem. The difference between used data is based to the different objective of the problems.
Berry et al. [15] describe how to place sensors into the edges of a water distribution system in order to minimize the population affected to a contamination event. The inputs for the algorithm are:

1. The network topology of the Water distribution system, is represented by a graph $G = (V, E)$.
2. $\alpha_{ip}$ is the probability of a contamination attack at node $v_i$ during the flow pattern $p$. For hypothesis it is possible exactly one contamination on a node during some flow pattern (only one node in which contamination starts).
3. $\delta_{ip}$, the amount of population who have access to the water distribution system at node $v_i$, during flow pattern $p$ is active.
4. $f_{ijp}$ is an index that takes two possible values, 1 if vertices $v_i$ and $v_j$ are connected inside a flow pattern $p$, 0 otherwise.
5. $S_{\text{max}}$ the number of available sensors.

The decision variable $s_{ij}$, is 1 if in the edge $(i, j)$ is placed a sensor, 0 otherwise. Given a contamination on node $v_i$ during flow pattern $p$, a node $v_j$ is contaminated only if exists a path between $v_i$ and $v_j$. Variable $c_{ipj}$ is used to denote that node $v_j$ is contaminated by node $v_i$ during $p$ is active. The mathematical formulation of the problem is:

$$\min \sum_{i=0}^{n} \sum_{p=0}^{P} \sum_{j=0}^{n} \alpha_{ip} c_{ipj} \delta_{jp}$$

s.t.

$$c_{ip} = 1 \quad \forall i = 0, 1, \ldots, n$$
$$s_{ij} = s_{ji} \quad \forall i = 0, 1, \ldots, n - 1, i < j$$
$$c_{ipj} \geq c_{ipk} - s_{kj}$$
$$\sum_{(i,j) \in E} s_{ij} \leq S_{\text{max}}$$
$$s_{ij} \in 0, 1 \quad \forall (i, j) \in E$$

The first set of constraints requires that when a node is directly attacked, it is contaminated, the second indicates that if a sensor is placed inside pipe $(i, j)$, then the pipe is monitored for both the flow direction, the third ensures that no maximum number of sensor placed is $S_{\text{max}}$, and the last requires that all the decision variables $s_{ij}$ may be binary variables. The objective function emulate the amount of people contaminated with the product $c_{ipj} \delta$, the multiplication for the variable $\alpha_{ij}$ require to give more importance at the most likely events. Limitations of the use of this methods are the fact that may be difficult to give a correct value of the variable $\alpha_{ij}$, and that the density population of an area may change dramatically during the period of time in which the network has to work. This approach does not fit with the realization of a wireless sensor network also because does not take into account any distance constraint between sensors, or any probability of error in the detection part.
A completely different approach for solve the sensor placement problem was provided by Byoung Ho Lee and Rolf A. Deininger [17]. In this work the objective of maximization is the amount of monitored water. Given a water distribution system $G=(V,E)$ and a node $v_i \in V$. Each neighbor node $v_i$ provides to it a percentage of the quantity of water present in each moment inside it. Placing a sensor in $v_i$ is therefore possible to monitor a portion of the water that was present inside its neighbors previously. In this work is built a coverage matrix in which is expressed this property for each node of the network. After placing a threshold value, a node $v_j$ is covered by a sensor placed in $v_i$ only if the percentage of water provided by $v_j$ to $v_i$ is greater than the threshold $t$.

At last Avi Ostfeld, and Elad Saloons work [16] propose to try to maximize not the coverage of the amount of water but of the possible events that can occur inside a water distribution system. Inputs of this problem are the topology of the water distribution system, the water demand and the other information about how water flows inside the system. After that it is builded an $n \times m$ random contamination matrix in which are defined $m$ contamination events and it is verified the probability of detection of these events in each of the $n$ nodes of the network. In order to solve the problem they use a genetic algorithm approach with probability of crossover of $p_c = 0.95$, mutation of $p_m = 0.02$, number of tested generation $n_g = 50$, the solution provided it is tested with EPANET, software used also in this thesis as simulator of water distribution systems. Also in this case it is not considered as a constraint the distance between the sensing units and the probability of detection of a sensor inside a specific node.

As described in the section above the detection in water distribution nodes have a lot of environmental and technical constraints, without taking into account them, the real performances of the network may not be satisfying. In fact it is possible to place sensor to cover the greater amount of water, or limiting the effect of a contamination but if in this node errors of detection may be frequent, the advantage of this placement is canceled inability of detection of bacteria in that node. At last in all these methods no constraint regarding the distance between nodes is required. The solutions provided are not topologies of wireless sensor networks, but is more similar to a set of independent nodes that cannot communicate each other.
Chapter 5

Proposed Optimal Sensor Placement Algorithm

In this chapter, an optimization problem is posed to find the placement of the sensors to monitor the water distribution system. The approach proposed is one of the possible. More solution methods are possible. For a general overview of optimization over wireless sensor network, see [21]

5.1 Optimization Problem under economic constraints

Given a water distribution system, the goal of this algorithm is to find the most appropriate sensors layout in order to maximize the probability of detection of the system. To realize that, is needed to take into account three points:

1. The probability of detection at each node is defined as discussed in the section 4.2 and it vary from node to node.

2. The maximum number of sensors is $S_{max}$, and is defined by economical constrains.

3. The topology of the wireless sensor network must be a connected graph. That means between each couple of nodes of the network that must exist a path.

Given a fixed economic budget, the number of array of sensors that it is possible to put inside the network is limited. The main goal of the optimization problem is to find the set of nodes in which put sensors in order to maximize the detection of the system, inside the entire water distribution network. A preliminary consideration is that more sensors are available, more the water distribution system is controlled by the wireless sensors network. It is possible to express introducing the function $\Omega$ defined as

$$\Omega = \sum_{i=1}^{n} p_{di} s_{i}$$
where \( p_{di} \) is the probability of detection in the \( i \)-th node and has always values greater or equal than zero, \( s_i \) is equal to one only if a sensor is placed in the node \( v_i \), zero otherwise. If sensors is placed always in the proper sites, the value of \( \Omega \) increases with the number of used sensors.

Consider the following definitions:

- \( V = (v_1, v_2, v_3, \ldots v_n) \) a vector with dimension \( n \) that includes all the nodes of the water distribution system.
- \( P_d = (p_{d1}, p_{d2}, \ldots p_{dn}) \) a vector with dimension \( n \) including all the associated probabilities of detection at each node \( v_i \).
- \( S = (s_1, s_2, \ldots, s_n) \) the vector that collect all the \( n \) \( s_i \) values defined above.

Then the function \( \Omega \) is

\[
\Omega = P_d^T S
\]

In case, find an optimal sensor placement in a water distribution system consists into find the value that maximize \( \Omega \), under the constraint

\[
\sum_{i=1}^{n} s_i \leq S_{\text{max}}
\]

that express that the maximum number of possible sensors must be lower or equal than the value \( S_{\text{max}} \). The formal formulation of this optimization problem is:

\[
\max_S P_d^T S \quad \text{s. t.} \quad \sum_{i=1}^{n} s_i \leq S_{\text{max}}.
\]  

Without other constrains, the proposed optimization problem returns the positions of the first \( S_{\text{max}} \) nodes with higher probability of detection. That because the product \( p_{d_i}s_i \) is always greater than zero. Anyway, a network based only on the information of the previous solution may have communication problems between sensors, because the distance between them is not considered a constraint.

To avoid this problem, it is required that the graph of the wireless sensor network, obtained by \( S \) as set of nodes, must be a connected graph. In graph theory given \( G = (V, E) \) two vertices \( v_i \) and \( v_j \) are called connected if \( G \) contains a path from \( v_i \) to \( v_j \). A path between two vertices \( v_i \) to \( v_j \) is a sequence \( v_i, v_k, \ldots, v_j \) of vertices such that from every vertex there is an edge to the next vertex in the sequence. A graph \( G = (V, E) \) is called connected if and only if all couples of vertices of \( V \) is connected, otherwise it is called disconnected. A connected component of a graph is a partition of vertices and edges of \( G \) such that the partition represents a connected graph; vertices of a partition do not share any path with vertices outside that path. For definition of connection, into a connected graph there exists only one connected component, otherwise a disconnected graph formed by two or more connected components. At last, the maximal connected component of a graph \( G \) is the connected component with the highest number of vertices. An easy way to test connectivity of a graph is to
run a search algorithm, such as breadth-first search algorithm, into it. Another way to establish if a graph is connected is to use some property of the adjacency matrix of the graph. The adjacency matrix $A$ of a graph $G$ is an $n \times n$ matrix in which the element $a_{ij}$ can takes two values:

$$a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

A fast connection test is this: given a matrix $A$, then if all the elements of the matrix

$$W = (I + A)^{n-1}$$

is greater than zero, then the graph obtained by using this matrix as adjacency matrix is a connected graph [22]. In the next section is described how to use the connection and how this constraint modifies the optimization problem is described.

5.2 Network connection test

Before inserting the constraint on the connection into the problem, some consideration must be done. Let is $G = (V, E)$, the topology of the water distribution system and $D$ an $n \times n$ matrix in which each element $d_{ij}$ take values:

- $d_{ij} = 0$ if the distance between nodes $v_i$ and $v_j$ is greater than a threshold $t$
- $d_{ij} = 1$ otherwise.

If $t$ is the maximum allowed distance between two wireless sensor devices, the matrix $D$ is the adjacency matrix of the graph $G'$ that has the same set of vertex of $G$, but have as edges the couples $(v_i, v_j)$, at physical distance lower than $t$. Let’s define $G'=(V,E')$ in which

$$E' = \{(v_i, v_j)|v_i, v_j \in V, d_{ij} \leq t\}$$

as the Communication Graph associated to the water distribution system graph $G$. The optimal sensor placement inside the network depends on the connectivity of $G'$. If $G'$ is connected, all nodes can communicate each other, so the region in which place sensors is the entire water distribution system; moreover the number of placed sensors is exactly $S_{max}$.

If it is disconnected, place $S_{max}$ sensors into the network is not possible. As explained in the Section 5.1, a disconnected graph is divided into two or more connected components. If inside a $G'$ there exists one or more connected components with a number of nodes greater than $S_{max}$, then it is possible to place $S_{max}$ sensors inside one of them, otherwise the sensor placing may produce always disconnected networks. In this last case the maximum number of sensors that is possible to place is given to the cardinality of the set of nodes of the maximal connected component of the graph. Based on this, it is possible to give this proposition:
Proposition 1. The maximum of sensors that is possible to place into a water distribution system $G=(V,E)$, is the number of nodes of the maximal connected component of the graph $G'=(V,E')$ where

$$E' = \{ (v_i,v_j) | v_i, v_j \in V, d_{ij} \leq t \}$$

Proof. Let be $k$ the cardinality of the maximal connected component of the graph $G'$. If it is possible to place $k+1$ sensors in the network, this means that in $G'$ there are $k+1$ nodes connected, in contradiction with the hypothesis.

Given $S^*$ optimal solution of the problem 5.1.1, using the information contained in $D$ it is possible to find the topology of the Wireless Sensor Network graph $G_1 = (V_1, E_1)$ in which

$$V_1 = \{ v_i | v_i \in V, s_i = 1 \}$$

$$E_1 = \{ (v_i,v_j) | v_i, v_j \in V_1, d_{ij} \leq t \}$$

Given the optimal solution $S^*$ and the matrix of distances $D$, the matrix of adjacency $A$ of $G_1$ is obtained by taking into account only columns and rows of $D$ that are referred to nodes with sensors are placed. A $i$th-row (or columns) is placed into $A$ if and only if component $s_i$ of $S^*$ is equal to 1. After computing $A$ is it possible to find the matrix $W$ as described in the equation 5.1.2. Than if all the elements of $W$ is greater than zero, the optimal solution $S^*$ provides a connected graph into which communication between nodes it is possible; otherwise if at least an element $w_{ij}$ is equal to zero, that means that the provided graph is disconnected. In this last case it is required to run again an optimization problem forcing it to avoid the previous solution. Let is $S_1$ a discarded solution at the first step, $k_1$ the amount of sensors of $S_1$ and $C_1$ an array defined as

$$C_1 = S_1^T$$

it easy to verify that

$$C_1S_1 = k_1$$

In the following solution of the problem $S_2$, in order to avoid the previous combination of $s_i$ elements, the subsequent inequality must be verified

$$C_1S_2 < k_1$$

Based on this, in order to discard all past solutions, the $i$th-optimization problem must verify the constraint:

$$CS_i < K$$

where :

$$C = [C_1, C_2, ... C_{i-1}]^T$$

is an $i-1 \times n$ matrix in which each row is a previous discarded solution, and

$$K = [k_1, k_2, ... k_{i-1}]^T$$

is an array in which the $j$th-element is the value $k_j$.

We are now in the position to describe how the proposed optimization algorithm works. At the first step the algorithm compute the solution of the optimization problem

38
Figure 5.1: Flow chart of the proposed algorithm

\[
\begin{align*}
\max_S P_d S \\
\text{s. t.} \quad & \sum_{i=1}^{n} s_i \leq S_{\text{max}}.
\end{align*}
\]  

(5.2.1)

This problem finds the first \(S_{\text{max}}\) nodes with the highest probability of detection. After that it tests the connection skill of the network provided by \(S\). If the network can communicate, then the algorithm returns \(S\) as solution of the problem, else iteratively computes the values of \(C\) and \(K\) described above and runs the new optimization problem

\[
\begin{align*}
\max_S P_d^T S \\
\text{s. t.} \quad & \sum_{i=1}^{n} s_i \leq S_{\text{max}}. \\
& CS < K
\end{align*}
\]  

(5.2.2)
until connected network is provided, or it is founded that the problem it is unfeasible.

The second optimization problem tries to find a combination of $S_{\text{max}}$ nodes with highest probability of detection avoiding as solution all the combinations founded before. Because optimization problem find always a solution that maximize $\Omega$, when the algorithm ends the founded solution is the combination of $S_{\text{max}}$ sensors with highest probability of detection that are able to communicate each other. The way of work of the algorithm is explained in the flow chart 5.1.

At last it is possible to describe a feasibility apriori condition of the problem. According to the Proposition 1, the maximum number of sensors that is possible to place into a graph is given by the cardinality $k$ of the maximal connected component of the associated graph $G'$. According to the first constraint there are two possibilities:

- If $k \geq S_{\text{max}}$, then the problem is feasible and the number of placed sensors is $S_{\text{max}}$
- If $k < S_{\text{max}}$, according to the first constraint, the problem is feasible and the number of placed sensors is $k$

So the feasibility condition of the problem consists in verifying that into the water distribution system it is possible to create a network with the lower values of sensors possible. The subsequent proposition follows the last observations:

Proposition 2. A necessary and sufficient condition to have a wireless sensor network placed into a water distribution system is that the maximal connected component of the associated communication graph $G'=(V,E')$ have a number of sensors greater than 2.

Proof. Let $G'=(V,E')$ be the communication graph associated to a water distribution system $G=(V,E)$. If more than two sensors belongs to the maximal connected component of $G'$, then it is possible to place sensors into these nodes and provide a simple connected wireless sensor network. Otherwise, if into a water distribution system a wireless sensor network is placed, the number of nodes of $G'$ that can communicate each other is at least equal to two. That means that in $G'$ it is possible to find a connected component with at least two nodes.

5.3 Proposed Algorithm

In this section the proposed algorithm for the sensors placement is described in pseudo-code.

Listing 5.1: Optimal Sensor Placement

```python
function [int Smax, int [] S] = Opt_Sensor_Pl(int [] V, int [] P, int Smax, int Dmax)
{
    // obtain the matrix of distances D
```

```
\[ D = \text{GetD}(V, D_{\text{max}}, S_{\text{max}}); \]

// obtain the connected components of \( G' \)
\[ [\text{Ncomp}, C] = \text{graphconncomp(sparse(D))}; \]

// obtain the connected components cardinality value
\[ \text{Cards} = \text{getCardinalities}(C); \]
\[ \text{optimal}_{S_{\text{max}}} = \max(\text{Cards}); \]

// check feasibility
if (\text{optimal}_{S_{\text{max}}} < 2)
  \text{return} \[ [0, 0, 0, \ldots, 0]; \]
else
  \{
    \text{Sum} = [1, 1, 1, \ldots, 1];
    // first version of \text{bintprog}()
    S = \text{bintprog}(-P, \text{Sum}, S_{\text{max}});
    \text{if (testconnect}(S, D, S_{\text{max}}>0))
      \text{return} S;
    \text{else}
    \{
      i = 1;
      \text{finded} = 0;
      C(i) = S';
      \text{while} (\text{finded} < 1)
      \{
        // second version of \text{bintprog}()
        S = \text{bintprog}(-P, C, c_{\text{val}}, \text{Sum}, S_{\text{max}});
        \text{if (testconnect}(S, D, S_{\text{max}}>0))
          \text{finded} = 1;
        \text{else}
        \{
          i = i + 1;
          C(i) = S';
        }\}
    \}
  \}
This algorithm takes as input a vector $V$ that contains the coordinates of the vertices, the value of $S_{\text{max}}$ defined as described in the section 5.1 and the distance threshold $D_{\text{max}}$. After that it computes the value of the matrix of the distances $D$ and using this obtains the connected components of $G'$ using the matlab routine $\text{graphconncomp}(\text{sparse}(D))$. Afterwards, the feasibility condition 2 is tested on the obtained graph, if the algorithm finds that the problem is infeasible, it returns immediately a null vector as solution, else it begins to find the optimal solution $S$. The algorithm tries to find the solution with another Matlab routine $\text{bintprog}()$ in two versions, the first using only $S_{\text{max}}$ as constraint, the second using also the matrix of discarded solution $C$. At last, every founded solution $S$ is tested by the following connection test:

Listing 5.2: Connection Test Function

```matlab
function [found] = test_connect(S,D,Smax)
{
    A=ComputeA();
    W=getW(A,Smax);
    for i=1:m
        for j=1:m
            if (W(i,j) < 1)
                return 0;
            end
        end
    end
    return 1;
}
```

This function test the connectivity of the provided network by computing the matrix of adjacency $A$ and the associated matrix $W$ defined in 5.1.2. Then it tests if all the elements of $W$ are greater than zero. If the test is positive, the function returns the value 1, 0 otherwise.

### 5.4 Computational complexity and speed up solution

In order to estimate the computational complexity of the algorithm it is possible to divide it into three parts:

1. Compute the matrix of distances $D$.

2. Obtain the connected components, and the maximum values of connected nodes.

3. Find the optimal solution $S$. 


The first part of the algorithm consists on the estimation of the matrix D. Given the vector V of dimension \( n \), compute D means that for each node must be estimated the distance between the others \( n - 1 \) nodes, that means this part has a computational complexity of order \( O(n^2) \). However it is possible to reduce the amount of distance calculated because the matrix D is a symmetric matrix, so

\[
d_{ij} = d_{ji}
\]

In this case the computational complexity is \( O(n) \).

The complexity of the second part depends on the complexity of the Matlab function \( \text{graphconncomp}(G) \). It takes a sparse matrix G and compute all the connected components of the graph obtained using G as adjacency matrix. Its computational complexity is \( O(n + e) \), where \( n \) is the number of nodes and \( e \) is the number of edges of the graph.

At last it is required to investigate the computational complexity of the last part of the algorithm. This part of the algorithm extract at each step a possible combination of \( S_{\text{max}} \) over \( n \) possible solution and test it. The test inside the cycle have a computational complexity of \( O(S_{\text{max}}^2) \) because only \( S_{\text{max}} \) values of V is tested. In the best case analysis the optimal solution is provided as the first of the possible combination, the computational complexity in that case depends only to the complexity of the test, \( O(S_{\text{max}}^2) \). In the worst case analysis the correct solution is the last combination of \( S_{\text{max}} \) nodes over \( n \). In that case the problem have a combinatorial explosion and the asymptotically computational complexity is

\[
O\left(\frac{n}{S_{\text{max}}}\right)
\]

. Using the proposition 1. So the global computational complexity of the algorithm depends on this part and is

\[
O\left(\frac{n}{S_{\text{max}}}\right)
\]

It is possible to implement a speed-up solution of the algorithm by taking into account for the possible combination only the nodes that belongs at connected component with more than \( S_{\text{max}} \) nodes. In that case the computational complexity in the worst case is

\[
O\left(\frac{k}{S_{\text{max}}}\right)
\]

, where \( k \) is the number of sensors that belongs at connected components with cardinality greater or equal than \( S_{\text{max}} \).
Chapter 6

Testing the solution

In this chapter we present the results of the optimization problem and of the simulations. In the first part we describe the topology of the network used for test the proposed algorithm. These topologies are provided by the software EPANET that is a simulator for the construction and the management of water distribution systems. Then we describe the other inputs of the optimization problem, such as the values that we use as probabilities of detection. After that we show the topologies of the wireless sensor networks provided by the algorithm. In the last part we show the result of the contamination scenarios realized with the software EPANET.

6.1 Network topologies

All the network topologies used in the simulation are provided by the water distribution system simulator EPANET. This software models water distribution piping systems, and provides also an integrated computer environment for editing network input data, running hydraulic and water quality simulations, and viewing the results in a variety of formats. These include color-coded network maps, data tables, time series graphs, and contour plots. Pipe networks consist of pipes, nodes (pipe junctions), pumps, valves, and storage tanks or reservoirs. EPANET tracks the flow of water in each pipe, the pressure at each node, the height of the water in each tank, and the concentration of a chemical species throughout the network during a simulation period. Chemical species, water age, source, and tracing can be simulated. We run three simulation for each topology, one in which non sensor are placed, the second where sensor are placed according to the optimal layout provided by our algorithm, the third one in which sensors are placed randomly. In this section are shown the networks used in the tests, the communication threshold and the connection properties of the communication graph associated at each topology. The amount of bacteria used for each scenario, the water quality thresholds and the values probabilities of detection inside each node are described further.

The first network used is the “Net1” of the EPANET software. This network has 11 nodes, 1 Pump, 15 pipes. The distance threshold used in this case is $D_{\text{max}} = 30$. The communication graph $G'$ associated to this network with this distance limit is a connected graph. The number of sensors $S_{\text{max}}$ that we place
in this scenario is $S_{max} = 6$. Since $G'$ is a connected graph and according to the proposition 2, the optimal sensor placement problem is feasible in this network.

The second network used is a network with 36 nodes, and 38 pipes. The used distance threshold is $D_{max} = 8$ and the communication graph $G'$ associated is a disconnected graph, in which the maximal connected component has 16 nodes. The number of available sensors in that case is 10.
Figure 6.2: Topology of the EPANET Example Network 2

Figure 6.3: Topology of the graph G’ associated to the EPANET Example Network 2

The last network used is the EPANET network example "Net3"; it is a very complex network, with 95 nodes divided as:

- 2 water reservoir (1 lake and 1 river)
• 3 tanks

The amount of edges in this network is 120. In the scenarios is used a communication threshold \( D_{\text{max}} = 3 \). This threshold provides a disconnected graph with a maximal connected component of 74 nodes, in the tested scenario the amount of sensors is 40.

Figure 6.4: Topology of the graph \( G' \) associated to the EPANET Example Network 3

In the next section it is explained the rest of parameters and inputs of the simulations.

### 6.2 Simulations details and optimal sensor placement results

In this section we describe the other values used inside them and inside the optimal sensor placement algorithm, it is described also the solutions provided by the optimal sensor placement algorithm. The most important is the array \( P_d \) that contains all the probability of detection \( p_{di} \) inside each node. As described in the section 4.2, it is impossible to have a measure of the probability of detection inside a node without have some test and measurements inside it. Therefore, in the simulation it is assumed that these values is available and it is considered as an input of the problem. Tables below express the probability of detection at each node of the three networks described in the section 6.1.

As described above, in the first example it is required to place 6 sensors inside the water distribution system 6.5, the amount of the probability of detection at
Table 6.1: Probabilities of detection of the first topology

<table>
<thead>
<tr>
<th>Node</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>0.3</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>0.7</td>
</tr>
<tr>
<td>11</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 6.2: Probabilities of detection of the second topology

<table>
<thead>
<tr>
<th>Node</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
</tr>
<tr>
<td>6</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>0.4</td>
</tr>
<tr>
<td>9</td>
<td>0.7</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
</tr>
<tr>
<td>11</td>
<td>0.6</td>
</tr>
<tr>
<td>12</td>
<td>0.9</td>
</tr>
<tr>
<td>13</td>
<td>0.5</td>
</tr>
<tr>
<td>14</td>
<td>0.8</td>
</tr>
<tr>
<td>15</td>
<td>0.4</td>
</tr>
<tr>
<td>16</td>
<td>0.5</td>
</tr>
<tr>
<td>17</td>
<td>0.3</td>
</tr>
<tr>
<td>18</td>
<td>0.4</td>
</tr>
<tr>
<td>19</td>
<td>0.3</td>
</tr>
<tr>
<td>20</td>
<td>0.8</td>
</tr>
<tr>
<td>21</td>
<td>0.7</td>
</tr>
<tr>
<td>22</td>
<td>0.7</td>
</tr>
<tr>
<td>23</td>
<td>0.7</td>
</tr>
<tr>
<td>24</td>
<td>0.3</td>
</tr>
<tr>
<td>25</td>
<td>0.6</td>
</tr>
<tr>
<td>26</td>
<td>0.8</td>
</tr>
<tr>
<td>27</td>
<td>0.3</td>
</tr>
<tr>
<td>28</td>
<td>0.5</td>
</tr>
<tr>
<td>29</td>
<td>0.2</td>
</tr>
<tr>
<td>30</td>
<td>0.4</td>
</tr>
<tr>
<td>31</td>
<td>0.3</td>
</tr>
<tr>
<td>32</td>
<td>0.6</td>
</tr>
<tr>
<td>33</td>
<td>0.1</td>
</tr>
<tr>
<td>34</td>
<td>0.5</td>
</tr>
<tr>
<td>35</td>
<td>0.6</td>
</tr>
<tr>
<td>36</td>
<td>0.4</td>
</tr>
<tr>
<td>37</td>
<td>0.8</td>
</tr>
</tbody>
</table>
Table 6.3: Probabilities of detection of the third topology

<table>
<thead>
<tr>
<th>Node</th>
<th>P</th>
<th>Node</th>
<th>P</th>
<th>Node</th>
<th>P</th>
<th>Node</th>
<th>P</th>
<th>Node</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7</td>
<td>21</td>
<td>0.7</td>
<td>41</td>
<td>0.1</td>
<td>61</td>
<td>0.8</td>
<td>81</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>22</td>
<td>0.5</td>
<td>42</td>
<td>0.4</td>
<td>62</td>
<td>0.7</td>
<td>82</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>23</td>
<td>0.4</td>
<td>43</td>
<td>0.2</td>
<td>63</td>
<td>0.4</td>
<td>83</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>24</td>
<td>0.4</td>
<td>44</td>
<td>0.3</td>
<td>64</td>
<td>0.3</td>
<td>84</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>25</td>
<td>0.6</td>
<td>45</td>
<td>0</td>
<td>65</td>
<td>0.2</td>
<td>85</td>
<td>0.7</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
<td>26</td>
<td>0.7</td>
<td>46</td>
<td>0.4</td>
<td>66</td>
<td>0.5</td>
<td>86</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>0.4</td>
<td>27</td>
<td>0.3</td>
<td>47</td>
<td>0.3</td>
<td>67</td>
<td>0.3</td>
<td>87</td>
<td>0.3</td>
</tr>
<tr>
<td>8</td>
<td>0.3</td>
<td>28</td>
<td>0.2</td>
<td>48</td>
<td>0.2</td>
<td>68</td>
<td>0.4</td>
<td>88</td>
<td>0.2</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>29</td>
<td>0.3</td>
<td>49</td>
<td>0.7</td>
<td>69</td>
<td>0.7</td>
<td>89</td>
<td>0.1</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>30</td>
<td>0.5</td>
<td>50</td>
<td>0.8</td>
<td>70</td>
<td>0.5</td>
<td>90</td>
<td>0.4</td>
</tr>
<tr>
<td>11</td>
<td>0.3</td>
<td>31</td>
<td>0.3</td>
<td>51</td>
<td>0.4</td>
<td>71</td>
<td>0.4</td>
<td>91</td>
<td>0.2</td>
</tr>
<tr>
<td>12</td>
<td>0.5</td>
<td>32</td>
<td>0.2</td>
<td>52</td>
<td>0.3</td>
<td>72</td>
<td>0.4</td>
<td>92</td>
<td>0.3</td>
</tr>
<tr>
<td>13</td>
<td>0.8</td>
<td>33</td>
<td>0.7</td>
<td>53</td>
<td>0.4</td>
<td>73</td>
<td>0.6</td>
<td>93</td>
<td>0.4</td>
</tr>
<tr>
<td>14</td>
<td>0.7</td>
<td>34</td>
<td>0.3</td>
<td>54</td>
<td>0.6</td>
<td>74</td>
<td>0.7</td>
<td>94</td>
<td>0.4</td>
</tr>
<tr>
<td>15</td>
<td>0.4</td>
<td>35</td>
<td>0.2</td>
<td>55</td>
<td>0.4</td>
<td>75</td>
<td>0.3</td>
<td>95</td>
<td>0.3</td>
</tr>
<tr>
<td>16</td>
<td>0.3</td>
<td>36</td>
<td>0.4</td>
<td>56</td>
<td>0.3</td>
<td>76</td>
<td>0.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.2</td>
<td>37</td>
<td>0.7</td>
<td>57</td>
<td>0.8</td>
<td>77</td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.5</td>
<td>38</td>
<td>0.6</td>
<td>58</td>
<td>0.4</td>
<td>78</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.3</td>
<td>39</td>
<td>0.3</td>
<td>59</td>
<td>0.3</td>
<td>79</td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.4</td>
<td>40</td>
<td>0.2</td>
<td>60</td>
<td>0.5</td>
<td>80</td>
<td>0.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

each node is described in table 6.1. The first iteration of the algorithm report as solution the wireless sensor network described in figure below, a green node means that has sensors placed inside.

![Figure 6.5: Topology of the first solution found by the algorithm](image)

The value of $\Omega$ for this solution is $\Omega = 3.9$, but as is possible to see in the figure the graph provided with this solution is disconnected, so it is discarded. After finding a connected solution, the algorithm found 23 different values of $S$ with disconnected associated networks. The connected topology has a value of $\Omega$ equal to $\Omega = 3.7$ and is shown in the figure below.
Figure 6.6: Optimal sensor placement for the first topology

Figure 6.7 shows the variation of the values of $\Omega$ during the iterations of the algorithm for finding a connected solution.

Figure 6.7: Variations of the values of $\Omega$ during the iterations of the algorithm

In the second topologies the 10 sensors of the network it is placed as shows in the figure 6.8
The algorithm takes 23 iterations after finding a connected network, the value of $\Omega$ in this case is $\Omega = 6.4$. Figure 6.9 shows how the values of $\Omega$ during the iterations of the algorithm.

Figure 6.8: Optimal sensor placement for the first topology

Figure 6.9: Variations of the values of $\Omega$ during the iterations of the algorithm
In the third topology it is required to create a wireless sensor network of 40 sensors. In this case the first connected solution is provided at the first iteration, the values of $\Omega$ is $\Omega = 31$. Figure 6.10 describe how the sensors are placed:

![Figure 6.10: Variations of the values of $\Omega$ during the iterations of the algorithm](image)

After describing the simulations inside the software EPANET in this section it we describe some other useful parameters. The first is the threshold value beyond which the sensors must react. E.Coli is used as indicator for fecal contaminations, combining with other bacteria (Enterococcus or C.Perfringens). As threshold value we use the standard threshold of admissible fecal contamination in water. The specifications of World Health Organization regarding this value are very strict and require the total absence of fecal contamination indicators per 100 ml of water (0/100 ml). However, this value can vary from country to country due to the economic and technological capacity of the same. For example, in European nations this threshold is met, but for example in Africa or South America, the value of allowable bacteria in drinking water is higher due to the lack of technology for water treatment in these countries. Fecal contamination indicators in Brazil for example in which is involved E.coli varies from 0 up to 5.

That means that in a practical realization of a wireless sensor detection system for E.coli, it is required to investigate at which particular values the network must be reacting. In the simulations is chosen the standard WHO threshold.

The amount of bacteria for each contamination event takes the values of 100 and 10 for each 100 milliliters of water provided by the contamination source. An overview of the simulations scenarios is that inside a network distribution
system, a source provides at each hour a fixed amount of contaminant bacteria (10 or 100), at first it is studied what is the effect of contamination without sensors placed, then in a different scenario with sensors placed it is simulated the effect of the disconnection of the sensing nodes inside the same water distribution system of the previous. Other parameters are the demand of water at each node (that may change the amount of water that flows inside it in one hour), the pressure and the level of the water. These parameters is fixed and provided by the EPANET network examples.

6.3 Contamination scenarios

The first contamination scenario involves the network topologies 1, where the injection node is the node 10 and the amount of contamination is 10 bacteria for each milliliter of water. The starting hour of each scenario is the 12 AM. After one hour node 1 is damaged, one hour later nodes 2 and 3 are infected and after 7 hours the network is completely contaminated. Node 11 is affected by contamination but with an amount always lower than the threshold. Since in this scenario no sensor is placed inside nodes, is not be possible to react at the contamination, so the network remain damaged until the end of simulation.

Figure 6.11, shows a qualitative representation of the contamination. In nodes in red the amount of bacteria is greater than the used threshold, in the blue ones no bacteria are present inside the node, other colors represent levels of contamination lower than the threshold (node 11 in the figure). In the second scenario sensors are placed inside the network according to the positions

Figure 6.11: Qualitative representation of the contamination when it starts (a), and when it reaches all the nodes of the network (b)
provided by the optimal sensor placement algorithm. In that case when a contamination arrive in a sensing node, it can detect the contamination and, in order to avoid that contaminated water flows inside network, disconnect itself to the network, after that other sensing nodes can detect a dangerous value of contamination and react to it.

Figure 6.12: Differences in amount of contamination between scenario 1 and 2 at node 6

Figure 6.12 shows how the amount of contamination differs inside node 6. The overall quantity of bacteria in the nodes is lower in the second scenario, the threshold value is reached earlier in the scenario with sensors than the other one, because water flows differently in this scenario due the disconnection of the sensing nodes. Results in the second scenario show that the sensors placement avoid contaminations inside several nodes of the network. Figure 6.13 shows the differences between the states of the water distribution system after seven hours. In the first case the contamination reached all nodes of the network, in the second because the sensing nodes 1, 5, 6, 8, after 4 hours detect the contamination and disconnected themselves from the network, nodes 11, 4, 7, 9 will never be affected by a contamination.
Figure 6.13: Contamination after seven hours in the first scenario (a), and in the second (b)

In the following scenario is proposed the same simulation values but it is used a random sensor placement. In this case, sensors are placed as shown in figure 6.14

The values of $\Omega$ in that case is $\Omega = 3, 4$ and it is lower than the optimal values, since that leakages of performances due errors of detection is greater in
the random placement than the optimal. Results show that when sensor are placed randomly the overall performances of the detection system is worse than the ones obtained with the optimal placement. Figure 6.15 describes the average amount of bacteria that flows inside the network for each hour in the optimal sensor placement scenario, and in the random one. As it is possible to see, in the optimal sensor placement scenario, the amount of bacteria is lower.

![Figure 6.15: Average amount of bacteria in the optimal and in the random sensor placement scenario](image)

In the next simulation the source node injects inside the example network 2 an amount of 10 bacteria for each 100 milliliter of water. Contamination starts
at 8 AM, and after one hour reaches the majority of the nodes of the network, the amount of contamination is greater than the WHO threshold is detected in 17 nodes out of 36 nodes. The network should be divided in four parts as shown in figure, in that the contamination source is labeled as "Source" and belongs to the area 0.

Area 0 and a part of the area 1 have been contaminated after one hour. After seven hours without sensor placed the entire area 1 and partially area 2 and area 3 are contaminated. After seventeen hours area 0,1 are completely contaminated and also a big part of the others are contaminated.

Figure 6.16: Graphical resume of the scenario

By placing sensors as suggested by the proposed algorithm it is possible to avoid contamination inside the areas 2 and 3.
Locations of the sensors in the random placement scenario for this topology are shown in figure 6.18.

The amount of $\Omega$ for the random placement is $\Omega = 5.4$ and is lower than the value obtained with the solution provided by the algorithm ($\Omega = 6.4$). Simulations shows that with this topology the effect of contamination inside the
network is higher than the optimal, in fact contamination reach also the area 2 and area 3 is the only safe area. Moreover figure 6.20, shows that inside the network the average amount of bacteria is higher in the random scenario than in the optimal one.

![Figure 6.19: Contamination effect and subdivision of the Example Network 2](image1)

Figure 6.19: Contamination effect and subdivision of the Example Network 2

![Figure 6.20: Average amount of bacteria in the optimal and in the random sensor placement for the Example Network 2](image2)

Figure 6.20: Average amount of bacteria in the optimal and in the random sensor placement for the Example Network 2

In last simulations 100 bacteria for each milliliter of water is injected inside the network topology 3. In this network the are five possible sources of water, in the simulations the contaminated source is labeled as Source in the figure. Contaminated source injects water in the network for 15 hours. 6.21.

After 15 hours most of the network is contaminated. Since during the propagation of contaminant two water sources is affected by contamination, network
remains contaminated also after the closing of the initial contaminated source. Graph 6.22 shows the variation of contaminant inside 4 nodes of the network, in which it is possible to see the effect of the water injected by the other contaminated sources.

When sensors are placed directed by the algorithm, contamination is restricted to the area near the contamination source. In fact, the two sources previously contaminated and does not suffer damage therefore in the rest of the simulation will continue to inject clean water into the network. The graphs above depict how differs the amount of contaminant in two of the four nodes of the network. By the action of the detection system, two of them is never reached by the contamination.
Random topology is tested and compared to the solution provided by the algorithm. In this case the sensors are placed according to figure 6.24.

The contaminated area in this scenario is slightly more extended than the optimal, the value of $\Omega$ in this case is of 20.4 instead of 24, obtained with the solution provided by the algorithm.
Figure 6.25: Graphical representation of the contamination after seven hours in the first scenario (a), and in the second (b)
Chapter 7

Conclusions

It is now possible to do some observations about the thesis work, the results of the simulations and give some possible extensions of it. Initial motivation of this work was investigate on the possibility to realize a Wireless Network of Biosensors, for having a real time detection of bacteria in water. It was importnat also have a clear point of view about the biosensors technology.

During the development of the work we found that currently no biosensors are available for bulding the system, anyway several companies are planning to realize this kind of sensors, so this is possible to have the physical realization of the system only after the end of the production of this units. Main contribution at the initial problem is the mathematical formulation and the provided algorithm, that find the optimal placement of these sensors inside the water distribution system. The maximization of the probability of detection is crucial for our model because only in this way we have the best performances of the detection system. We propose an algorithm that provide the optimal solution by taking into account topological information and an estimation of the probability of detection inside each node. This algorithm has a not polinomial complexity, anyway this limitation is not a problem because is possible to run the algorithm when the system is off-line.

In the simulations part we have proved that the solution provided by the algorithm have the best performances in all the three network tested. We have seen that the solution provided by the algorithm preserve better water distribution system to contamination event. The values used for the simulations are real and actual thresholds or indices of water quality.

There are several possible extension of our work. At first is important to develop the software for data management. The software may be an helpful instrument for the identification of the contaminated nodes, the level of pollution, and for the maintenance of the network. In general this software may represent a user-friendly interface that permits to manage the network, with the intent to make the interaction with the detection system easy also for not particularly skilled personnel. Other extension is to use the algorithm for having some different results, for instance running the algorithm it may be possible to find the minimum number of sensors that permits to have a fixed values of $\Omega$. This is may seen as utility that provides an information about the economical convenience of the realization of the network. Under economical constraints, the realization of a system like the proposed may be difficult, so a preliminary test
that returns the number of sensors needed to obtain a minimum value of network performance may be helpful in deciding whether it is worth to realize the detection system. Moreover, if the given economic constraints permit to realize network with an higher number of sensors, it is possible to use this utility to decide to place a lower number of sensors, and allocate the rest of budget in other policies such as maintenance and security.
Bibliography


[22] Horn and Johnson, Matrix Analysis.