CHALMERS

Security Aspects in the Advanced Metering Infrastructure

Master of Science Thesis in Computer Science

MIHAI COSTACHE

VALENTIN TUDOR

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering

Göteborg, Sweden, December 2011

Security Aspects in the Advanced Metering Infrastructure

Master of Science Thesis in Computer Science

MIHAI COSTACHE

VALENTIN TUDOR

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, December 2011

Mihai COSTACHE,
Valentin TUDOR,

Examiner: Magnus ALMGREN
Supervisor: Marina PAPATRIANTAFILOU

Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden, December 2011

Abstract

The traditional electrical grid is transitioning into the *smart grid*. New equipment is being installed to simplify the process of monitoring and managing the grid, making the system more transparent to use but also introducing new security problems. Smart meters are replacing the traditional electrical utility meters, offering new functionalities such as remote reading of the consumption indexes, different time of use tariffs, automatic error reporting, and the possibility for the electricity providers to remotely turn off and on the electricity service at one location. This research thesis studies this last feature through two scenarios where we emphasize the effects of an attack exploiting the remote turn off feature, both on a theoretical level and through a simulation. In the first scenario, the *frequency* property of the grid is the target in an attempt to cause a widespread blackout. In the second scenario, the *voltage* is driven out of bounds by the adversary, causing physical damage to the electrical appliances of the affected customers.

Data provided by the smart meters can be used to develop fraud and attack detection and mitigation tools. Obtaining real data can sometimes be cumbersome, due to privacy concerns. We propose an anonymization technique for sensitive data, based on a cryptographic procedure; this provides consistent results even if it is used over different traces. An implementation of this technique is also provided. In the process of developing fraud and attack detection and mitigation techniques, the case of off-line centralized data is covered, for both individual smart meters and clusters of smart meters.

Acknowledgements

We would like to express our gratitude towards our advisers, Dr. Marina Papatriantafilou and Dr. Magnus Almgren for their valuable insights and constant feedback during the elaboration of this research thesis. They have encouraged us to explore additional domains in the quest of finding answers to the problems that appeared along the way.

We would also want to thank Dr. Christopher Saunders, from the Energy and Environment Department, for his useful inputs regarding issues from the Electrical Engineering domain.

Last but not least, we would like to thank our families for their continuous support and understanding.

Mihai Costache and Valentin Tudor, Göteborg, 2011/10/15

Contents

1

Introduction

The electrical distribution grid is transitioning from the traditional grid into the new so-called *smart grid*, partly to become more flexible and to be able to accommodate large energy production from renewable sources. This transition involves, among other steps, the installation of advanced equipment in places where previously it was not found, such as the distribution section of the electric network. The most important change in the distribution part of the electrical network is the introduction of *smart meters* replacing the traditional domestic electrical meters. Even though this transition enables new functionalities, it also brings security concerns regarding how the technology can be misused by a malicious adversary. Many of the new security issues in the smart grid are well-known problems in the information and communication technology (ICT) domain, such as buffer overflows in devices and sloppy implementations of cryptographic protocols, but some issues originate from the electrical and power engineering domain (device tampering). There are also new challenging problems, requiring an interdisciplinary approach for the analysis of possible solutions with regard to electrical grid and IT infrastructure simultaneously.

1.1 General information

Nowadays the primary globally-consumed resource to produce electrical energy is coal, which together with natural gas and oil account for 67% of the total energy produced in 2009 [1]. Nuclear power covers another 13%, while the main source of renewable energy comes from hydroelectric plants (16%). Solar, wind and geothermal energy cover only 3% of the energy production. However, the long-term strategy of many countries is to use more renewable energy production. Germany, for example, has recently announced that all nuclear plants operational from before 1980 will be shut down by the end of 2011, and by 2022 all the nuclear power production should be ceased [2]. The plan is to replace the nuclear energy with renewable energy, which by 2020 should count for 35% of the national energy production, i.e. double than what it is today, as well as to decrease the electricity consumption by 10%. The migration to use more renewable energy is one factor driving the adoption of the smart grid, together with the expected wide adoption of hybrid vehicles as well as a better utilization of produced energy, meaning that both the traditional energy transmission and distribution networks are being upgraded.

The migration towards using more renewable energy is one factor driving the adoption of the smart grid, but other driving factors are the expected wide adoption of hybrid vehicles as well as a better utilization of produced energy, meaning that both the traditional energy transmission and distribution networks are being upgraded.

As part of this process, the EU mandates that all the metering devices present in the traditional energy distribution network should be replaced with smart meters by 2020[1], in an attempt to better control and monitor the energy consumption aiming to create a cleaner environment to live in. Some countries have already completed the installation of smart meters in the distribution network, such as Sweden, Germany, Italy, and UK.

The *smart meter* allows remote reading of consumption, automating the collection of monthly consumption indexes by minimizing the need for an operator to manually read each meter. Another goal of smart meter's deployment is to influence the consumption behavior of end-users by providing near-real-time information on measured indexes. While the current definition of a smart grid is abstract, overall it can be summarized as "electricity networks that can intelligently integrate the behavior and actions of all users connected to it – generators, consumers and those that do both – in order to efficiently deliver sustainable, economic and secure electricity supplies" [3]. Simply put, the main purpose is to extend the traditional network so it becomes more flexible by adding new equipment and a management layer; this layer controls the equipment, making the system robust, flexible, easier to administer and more transparent. This change is necessary, among other reasons, to accommodate the use of more renewable energy sources. However, meters will also have other functions such as promptly alerting the distribution company of electrical problems occurring at the site of the customer (such as power outages), and maybe even control when and what devices are allowed to run.

From the ICT perspective, the smart meter has the ability to measure, process, and communicate with other meters, data concentrators, or the central system. As such, these integrated ICT components bring many new functionalities to the grid, but also lead to many security problems already present in traditional ICT systems, albeit with a big difference: the electricity network is a critical infrastructure in society and if it fails, many other systems will in turn cease to function correctly. Problems also stem from different underlying assumptions in the ICT domain compared to the electrical engineering (EE) domain. In electricity networks, equipments are expected to have an extended life span (about 20 years), while within the ICT domain it is not unusual to patch systems on a weekly basis. Replacing a large number of smart meters after they are installed, or simply updating their firmware might be very costly. For that reason, the systems must be planned well from the beginning with a good security model. Unfortunately, as explained later in the paper, several vulnerabilities in these

[1] http://ec.europa.eu/clima/policies/brief/eu/index_en.htm

systems have already been discovered.

The smart meter and other technological advances enable further changes to the electrical grid. From the centralized system with a few large energy producers, where energy is *broadcast* to the consumers, local renewable energy production through solar or wind power turns the grid into a more distributed structure [4], by creating local power generation areas called *power islands* or *micro grids*. Some islands then become self sufficient, and may even inject (sell) their surplus energy back into the distribution network. Keeping track of the energy consumed and the energy injected requires the installation of smart meters. But apart of the modifications that appear on the Information Technology part, the electrical network is also subjected to changes. The presence of more renewable energy production units (solar and wind power) in the premises of end-clients locations will make the electrical network shift from today's centralized structure to a more distributed structure.

1.2 Purpose and objectives

The communication between the smart meters and the Central System (Metering Data Management System) is made through channels which have been proved to be prone to security breaches [5]. Security flaws have been discovered in the current implemented smart meters. In [6], McLaughlin et al. present a scenario where injecting false malicious data leads to gaining different benefits from the system. So simply taking control over some smart meters in order to create havoc in the network is not an unachievable goal in a realistic scenario.

In this thesis, we present two scenarios where a similar type of distributed attack is used against smart meters but executed at two different scales. In the first scenario, we focus on the electrical distribution network from a large geographical region (several large cities) where the adversary tries to take control over a very large number of smart meters. By using the remote capability to turn the smart meter on or off, the adversary can tamper with the frequency of the electrical grid. The second scenario is localized to a neighborhood modelled as a power island. Here, the attacker's purpose is to create havoc and damage the electrical appliances in the neighborhood by changing the voltage in the network through his control over the smart meters. We stress the importance of consumption data to create a realistic simulation setup as well as for anomaly detection. With respect to privacy of the customer, we advocate that a data anonymization process is necessary, by means of cryptography-based approach similar to the work in [7]. Mitigation techniques can be developed by analysing consumption data from the management center, either at individual customer profiles level or at clustered level. For instance, an individual profile can set the base rules for energy theft detection by measuring abnormally low deviation from the the average behavior. On a large scale, clustering algorithms can prove to be more suitable in detecting consumption irregularities within a specific group.

1.3 Structure of the thesis

The rest of the thesis is organized as follows. Chapter 2 describes the features of smart meters and their communication infrastructure, as well as the electrical concepts that are necessary to understand the simulation of the second scenario. Chapter 3 presents the related work from different research areas relevant to Advanced Metering Infrastructure and describes how our study integrates in this new strand of research. In Chapter 4, we outline the two scenarios, the possible consequences and study the second scenario in further detail using a simulation setup. Chapter 5 covers consumption data privacy and analysis in the scope of devising mitigation techniques. Chapter 6 concludes this thesis and draws the outline for future research.

2

Background

FOR an easier understanding of the attack scenarios described in Chapter 4, we present a brief overview of the smart meter's main features and its communication model. We also include a short summary of important terms and formulas used for the simulation of the electrical network.

2.1 Smart meters

A smart meter is an embedded system whose main current functionality is to automate the collection of consumption indexes by minimizing the need for an operator to manually read each meter. Conceptually, it can be seen as having three components: the electrical meter, the processing unit and the communication module, as shown in Figure 2.1.

The electrical meter has the role of measuring the electricity consumption at the power line and to translate the readings into data that can be used by the processing unit. The processing unit's role is to process and store the information and to control both the electrical meter and the communication module. The communication module can be embedded in the smart meter or installed in an extension slot, meaning that the same smart meter equipment can use different communication modules depending on the

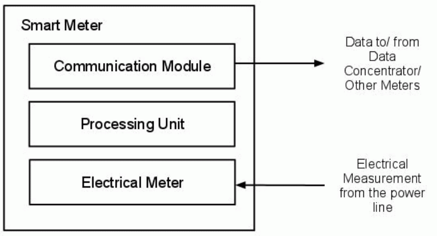
Smart Meter
Communication Module
Processing Unit
Electrical Meter
Data to/ from Data Concentrator/ Other Meters
Electrical Measurement from the power line

Figure 2.1: Smart meter components

circumstances. In some deployments, a ZigBee network is used in urban neighborhoods while GPRS is used in rural environments.

The electrical meter part of the smart meter is usually highly regulated and must conform to a set of national standards. The other two parts vary in their functionalities, but the more advanced ones may have an indoor module to inform the customer of their instantaneous electricity consumption, energy consumed so far (translated into a currency), or the current tariff based on time of day utilization. One key feature of the smart meter is the remote ON/OFF switch. The distribution company can, for example, cut the power from customers that have defaulted on their payments by simply issuing a remote command to the smart meter, without sending a crew to the premises to physically perform the disconnect procedure.

Real-time reporting of energy usage of the end-customers will also enable a better management of the distribution grid by reducing electricity loss and maintaining efficiency of electricity production. Smart meters have the capability to report information to the data concentrators at different intervals (daily/hourly), but this may change to real time reporting in the future. Some meters can be queried about the current load, raising some privacy concerns since such data can be used to identify what appliances are currently in use.

Finally, customers can choose to install renewable energy production facilities in the premises of their homes (such as solar panels) and some customers may then produce more energy than they consume. The smart meter is responsible for keeping track of the energy consumed from the electricity network as well as for the energy injected back into the grid.

2.2 Communication protocols

The smart meters can send data via different communication channels (IP, GSM, GPRS, PLC, ZigBee) to so-called *data concentrators*, to aggregate data and then dispatch these to the central system, the *metering data management system*, as shown in Figure 2.3.

2.2.1 DLMS/COSEM

As we have seen earlier, the smart meters need to transmit the metering data over to the data concentrator to be forwarded to the MDMC (Metering Data Management Center) where various operations are executed. One such operation that is of particular interest is the calculation of the cost of electricity consumed by each particular customer, which is then used for billing purposes. Because metering equipment has a special design and the data formats differ from the traditional IT communication, it needs a specific communication protocol for interfacing with the IT equipment already available. The proposed protocol took the name of DLMS/COSEM (Device Language Message Specifications/Companion Specification for Energy Metering), and is available as a standard. The communication model is similar to a client-sever architecture, with the difference that here the smart meter acts as a server, while the data concentrator acts as the client.

The data exchange occurs on demand, when the client (data concentrator) sends a request to the server (smart meter), or in trigger based manner when an exceptional or previously set event occurs. It also inherits the OSI layered protocol architecture, but has variations depending on the communication configuration used.

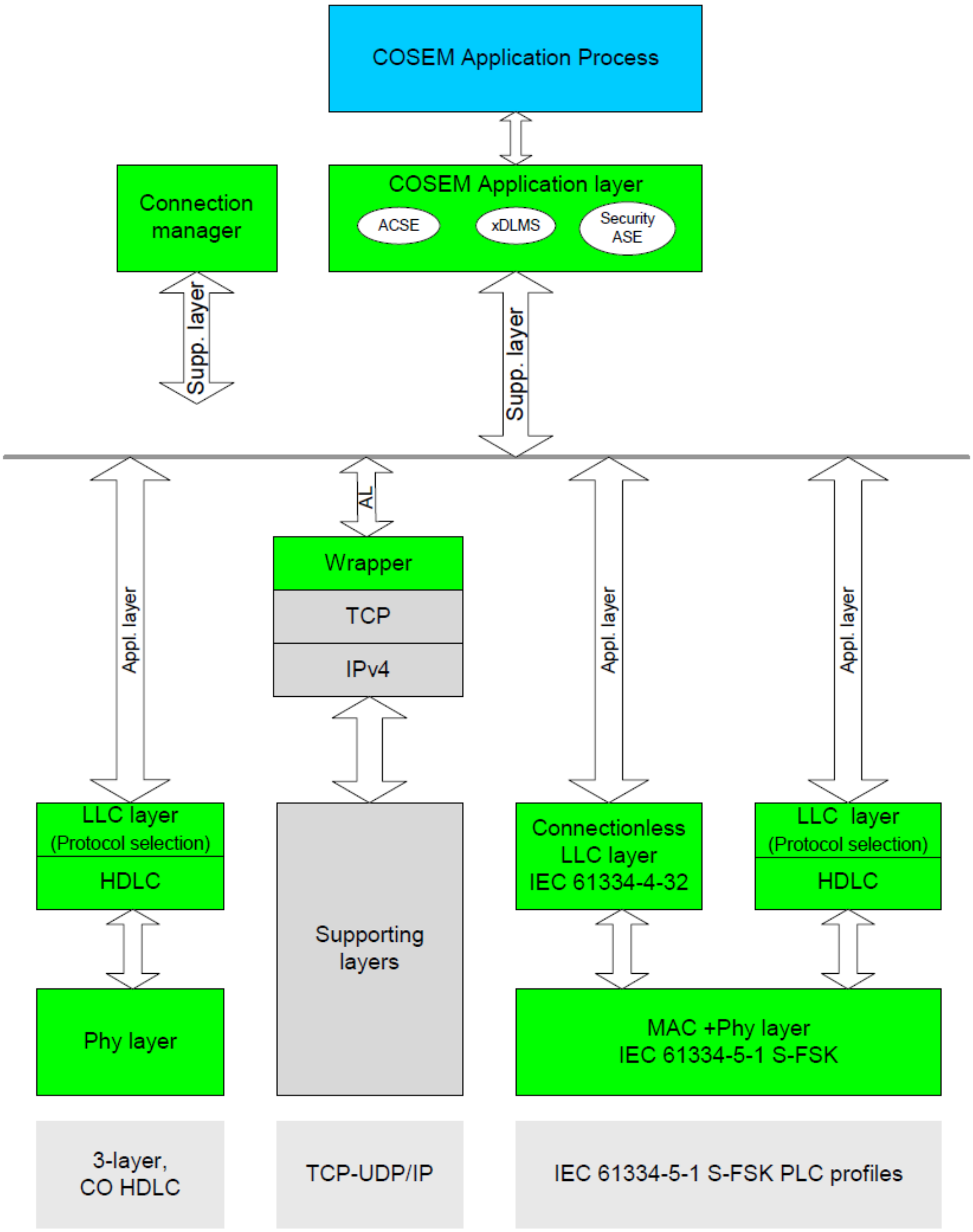

Figure 2.2: DLMS/COSEM Communication Profiles [8]

The variations in the protocol layer stem from the type of network infrastructure available or chosen by each energy provider. Thus, there are cases when energy companies opt for owning their IT infrastructure, so DLMS/COSEM must interface to the TCP or UDP communication equipment used. Other times, because owning an infrastructure is too expensive, interfacing with the protocols used by telephone networks (PSTN, GSM) is cheaper. Another cheap alternative is using the power lines themselves for data transmission (at the cost of a low data rate transmission), but this technology is limited in range, so it needs to be based on the previously mentioned methods to ensure that data reaches the management center. Figure 2.2 presents different communication protocols that are supported by the DLMS/COSEM architecture.

2.3 Security

In any new type of electronic device, security is one major concern. Because the smart meter incorporates modules for data conversion, storage and transmission, each of those modules are subject to vulnerabilities, so they need a configuration with security in mind. As we will see in the following sections there are different measures taken to ensure the safety of such devices.

2.3.1 Physical security

From a physical perspective, to ensure that the device is tamper free proves to be difficult. This is mostly because the smart meter is installed in a non-secure environment, at the premises of the customer. Some cases present a lower security threat, when the smart meter is installed inside the residence of the customer, so an attacker will not have free physical access and he must perform a physical intrusion, unless the attacker and the owner of the house are one and the same person. In other cases the smart meter is installed outside the residence of the customer, easily accessible to anyone. To prevent the physical tampering of the smart meter the old traditional seal is used. With the technological advancements a seal is now software bound, in the sense that if this seal is tampered, an exceptional event is recorded and thus an alert is transmitted to the meter management center. But one seal reporting is not enough as we have seen in [9] that a seal can be triggered by railway passing, and even then one such alert is in most cases ignored by the energy companies because it is questionable whether to disconnect or not a customer for such reason.

2.3.2 Software security

Every smart meter is able to run its functions because of the firmware which is installed on the hardware. Due to its novelty, there is no stable and vulnerability free version yet, a reason why periodical patch updates need to be considered. This needs to be viewed as a critical action and must be conducted with proper care. There are different methods of patching the firmware of the smart meters when a vulnerability is discovered. One method is to distribute patches from the center to each smart meter. Another interesting method is using the peer communication of the smart meters to propagate patch updates.

The problem with this approach is checking that all devices have been properly updated. But also this is a questionable approach since we have seen in the last section the security problems with physical tampering. A device that has been tampered with may be used to spread a malicious version of the firmware update. Standards such as [10] propose for each smart meter an approach that requires more computational power and memory in order to accommodate firmware diversity. This way each device retains two firmware versions, one of which is the updated version. Before the updated version is applied to the smart meter, it is checked for consistency and conformance, and this method is believed to solve spreading malicious updates described earlier.

2.3.3 Communication security

As described in section 2.2.1, there are various communication environments used to transfer data from the distribution side, where the customer and the smart meters reside, to the data concentrators and management center. Because those communication environments are adapted from the ICT world with security as a secondary objective, they also inherit the weaknesses present here. For example, the first draft versions of the DLMS/COSEM protocol had no security measures in place, such as message authentication, hash function capabilities or encryption; these were only added in the latest version [8]. With these communication configurations, different security approaches are implemented.

Power line communication uses physical layer signal modulation. One attempt, in [11], was the use of OFDM (orthogonal frequency division multiplexing), offering robustness, simplicity, performance, compatibility, scalability and security. Aggregating layers into the transport layer allows interoperability with IP protocols. Besides the signal modulation, the OFDM PLC profile uses meter authentication with white-listing, black-listing, CCM encryption and protected channel EAP-PSK.

For the case where the distribution company owns the communication infrastructure, for example in Gothenburg, Sweden, the grid is equipped with smart meters having wireless transmission capabilities with ZigBee chips [12]. Communication security in this case ensures authentication through digital certificates, data confidentiality with AES-128 encryption and non-repudiation of messages with digital signatures. There is ongoing research in the field of communication security aiming for a secure environment and dealing with current vulnerabilities, further described in Chapter 3.

2.4 Electrical Engineering

2.4.1 Transmission and distribution networks

While smart meter communication resides in the communication engineering area, aspects from the electrical engineering domain also present interest. In the context of rising demand for energy and the shift towards renewable sources, the grid's infrastructure limits are questioned.

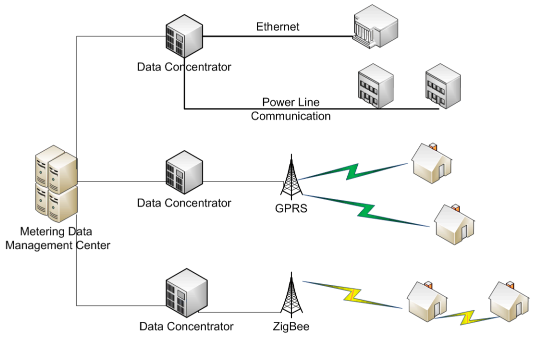

Figure 2.3: Smart meter communication model

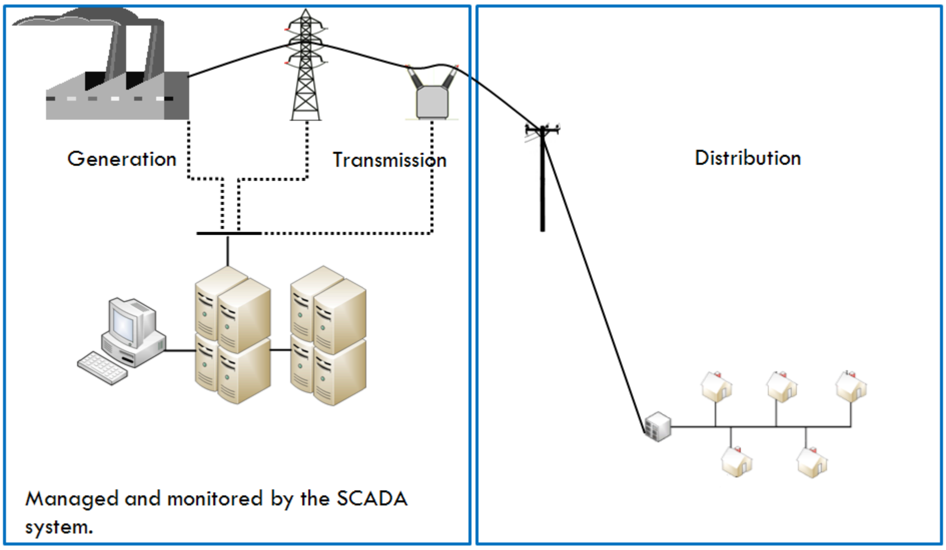

Figure 2.4: The Traditional Electrical Grid

Power generating facilities are usually placed in remote locations and therefore a transport network is necessary to deliver electricity the end consumers. A typical electric grid has two components: the *transmission section* and the *distribution section*. The transmission section transports the electricity from the generator to the distribution section and the distribution section delivers it to the end customers. Transporting over long distances requires electricity to be converted to a high voltage alternating current form (HVAC) that enables efficient delivery with an acceptable loss. This is done by using a step-up transformer that outputs voltages in the range of $50-350$kV (depending on the line capacity) into the transmission grid. Before reaching the end consumer, the transmission grid is connected to a series of step-down transformers that feed lower voltage electricity to the distribution grid (typically $0.4-50$kV), following a multistep transformation from high to medium and from medium to low voltage.

2.4.2 Underground and Overhead distribution lines

In the distribution part of the grid there are two types of lines that can be used with regard to the place where they are installed. If the distribution lines are mounted on poles (above the ground) they are called overhead distribution lines and if they are mounted in utility ducts (underground) they are called underground distribution lines.

Choosing between the two types of line mounting is done depending on the area that needs to be served. According to [13] there are two key-points on which the debate between the usage of underground vs overhead lines focuses: reliability and cost. Underground lines experience problems less often than overhead ones, but when a problem occurs, the time required to fix it is longer because the lines must be dug up in order to be inspected and repaired. The life expectancy of underground lines is shorter than the one of overhead lines, because the first ones are prone to natural processes such as rust or oxidation. The costs of installing underground distribution lines is about 4 to 6 times the cost of installing overhead distribution lines, according to the same document [13]. There is a strong opinion towards installing underground lines, especially in urban areas, for aesthetic reasons, but usually the extra costs reflect in increased rates for the end customers.

Choosing between one of these two distribution line types has an effect on the types of cables that will be installed. Pabla specifies in [14] that for overhead distribution lines are used ACSR (Aluminium Conductor Steel Reinforced) cables, while for underground distribution lines, aluminium cables are used. The main electrical difference between these two types of cables is that the inductance for the ACSR cable is about four times higher than the inductance for the Al cable. A higher inductance means that the line experiences a higher voltage drop, a property which is important in the attack simulation conducted in Section 4.4. The effects of the attack are expected to be more dramatic in the case of the overhead distribution line than in the case of the underground distribution line.

2.4.3 SCADA system

As described in the previous section, the electrical grid is divided into two parts: transmission and distribution. The transmission part of the grid is monitored and controlled by the Supervisory Control and Data Acquisition system or shortly SCADA. This is a computerized system that gathers data from sensors placed in the generation and the transmission part of the grid. Based on the information provided by these sensors and using a virtual model of the generation and transmission network, the actual state of the network can be inferred, one important component in the process of monitoring the system, as showed in [15]. The same model can be used to simulate hazard scenarios (such as power production interruption, line tripping or equipment malfunction) in order to create better solutions, if these scenarios would occur into real-life, or it can be used for economic purposes in order to simulate network load and optimal power routing.

The criticality of such a system can be derived from its functions. The SCADA system is used to ensure the well-function of the transmission network, being able to connect or disconnect high voltage power lines in order to redirect power on other ways to ensure the grid functions within its designed capacity. One of the most referred to incidents caused by problems in the SCADA system, is the 2003 blackout that affected the US and Canada due to failure in monitoring and managing active power reserves. Then 50 million people were left into darkness and the estimated loss was approximately 6 billion dollars.[1] Research on how to improve the security of such systems are multiple and references are presented in Chapter 3.

2.4.4 Power quality

One of the important factors in the design of a distribution grid is power quality, i.e. limits for power supply frequency and voltage magnitude, so that electric and electronic equipments can function without damage when connected to an outlet. It is important to account for the transmission line's loss to ensure power quality standards for each *feeder*, where a feeder is a portion of the grid that provides power transportation capabilities to service areas. In real-world conditions with variable loads and unpredictable power generation levels, power quality issues need to be handled in order to respect the specification of appliances. According to the European "Voltage Characteristics in Public Distribution Systems" including EN 50160 and EN 61000 standards [16], there are specific voltage requirements and frequency regulations for different situations. There are requirements for an acceptable variation of voltage magnitude (from $220 - 240$V nominal value, depending on country) and power frequency (50Hz nominal value). Variations allowed for the frequency must be on average $\pm 1\%$ ($49.5 - 50.5$Hz) in 95% of a week time and $-6\%/+4\%$ ($47 - 52$Hz) at all times. Voltage magnitude variation should be within $\pm 10\%$ of the nominal voltage in 95% of a week time and all average values should not go outside $-15\%/+10\%$ of the nominal voltage. In the case that an energy provider does not respect power quality standards, grid problems may appear and cascade to unforeseen consequences. Therefore, power quality specifications are enforced by financial penalties on the energy providers. Problems related to the quality of the voltage may manifest themselves as short interruptions, flickers, voltage dips, supply voltage variations and harmonic disturbances. For more information, we refer the reader to [17], which explains in detail how these phenomena manifest.

In the attack scenarios described in Chapter 4, the electrical power frequency or the voltage magnitude, respectively, is pushed outside of the standard value limits. By inducing abrupt variations in the loads at precise points in time, for example when the grid becomes under loaded (too much available power), the voltage magnitude can be pushed outside the range of safe values. Formally, this is a consequence of the power flow equations relating individual nodes or bus power properties, used for the simulation in Chapter 4.4. These equations are briefly outlined below.

[1] NYISO Interim Report August 14, 2003 Blackout `http://www.hks.harvard.edu/hepg/Papers/NYISO.blackout.report.8.Jan.04.pdf`

2.4.5 Power flow equations

The electrical grid is composed of *nodes* or *buses*. A node or a bus is a point in the system where different properties, such as voltage or power, can be measured and are used in electrical modelling [18]. Each node is connected by a *transmission line*. In order to model the electrical infrastructure and electricity flow, each transmission line is characterized by physical properties (resistance, capacity and inductance). The line's specific properties can be measured, and the loss characterizing the line, also known as impedance can be calculated. A square matrix, whose dimension depends on the number of nodes in the network, can be built to represent the impedance between nodes (the Z matrix or impedance matrix). However, the Y matrix (or admittance matrix), the inverse of the impedance matrix, is used in practice. With the following power flow equations, voltage magnitude and angle at each node, as well as real and reactive power flowing in and out of each node, can be computed as:

$$\begin{aligned} P_i = & \sum_{i=1}^{N} Y_{ij} V_i V_j \cos\left(\theta_{ij} + \delta_j - \delta_i\right), \\ Q_i = & - \sum_{i=1}^{N} Y_{ij} V_i V_j \sin\left(\theta_{ij} + \delta_j - \delta_i\right), \end{aligned}$$

where P_i and Q_i are the real and the reactive power at node i, respectively; Y_{ij} and θ_{ij} are the magnitude and angle of the admittance between node i and node j; V_i and V_j are the voltage magnitudes at node i and node j; δ_i and δ_j are the phase angles at node i and node j.

Based on these values, voltage regulations can be enforced for each feeder in the grid [19]. Grid control and regulation in a centralized system can be solved by central operators in the SCADA (Supervisory Control And Data Acquisition) system. However, in the context of distributed generation where customers can produce their own energy, and thus become providers for their neighbors, serious problems may arise; this is the actual topic explored in the second scenario in Chapter 4. For instance, since power injection into a grid is in some regions freely allowed (with the required standard specifications to be met), a node's voltage magnitude may vary even more due to abrupt changes in consumption.

2.4.6 Power islands

One of the main changes involved in the transition to the smart grid is the distributed generation of energy. In [20], distributed generation is defined as: "[...] an electric power source connected directly to the distribution network or on the customer side of the meter". Many customers will opt for installing a renewable energy production facility on their domain, for example a wind turbine or photovoltaic panels, in order to obtain some independence from their main energy provider; some may even sell the surplus energy produced.

The traditional distribution network can be modelled as a tree-like structure, but with the changes of local producers of energy the best model is more flat, like interconnected power islands. The second scenario (the voltage variation scenario) presented in

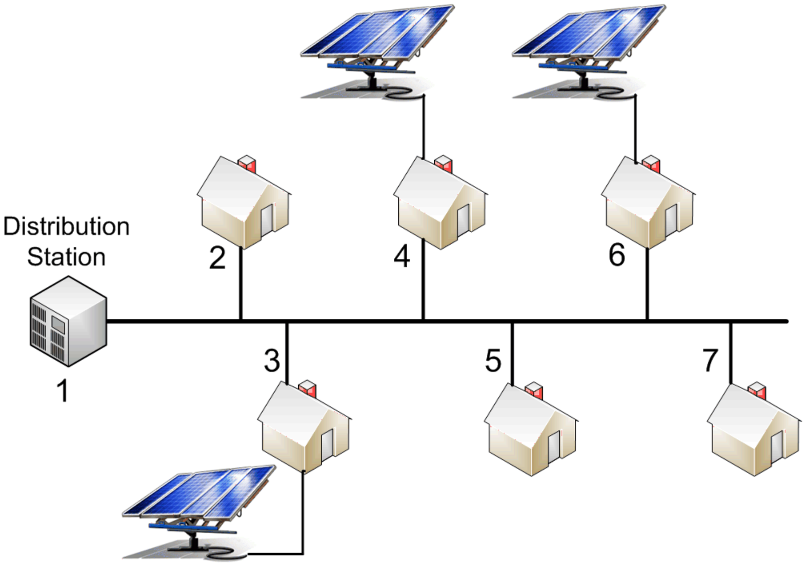

Figure 2.5: A power island with renewable energy sources

Chapter 4 takes place in such a power island (see Figure 2.5), which has become self sufficient and surplus energy is injected back into the grid. The power island is composed of six residences, all of them served by a electrical distribution station, and three of the residences have renewable energy sources installed on their premises.

3

Related work

The European Union's vision is that by 2020 all the members must achieve several goals regarding the electricity production and consumption process. In order to reduce the carbon footprint and the greenhouse gas emissions, the energy consumption should reduce by 20% and also 20% of the quantity of energy produced by a country must come from renewable energy sources.[1] To better monitor the energy consumption and to improve the ratio between the energy produced and the energy consumed, more detailed and complete information is required. This information must come from the energy end-users with a higher frequency and in order to obtain this information all the traditional electrical meters should be replaced by smart meters in all the European countries. This is an important step in the modernization of the traditional electrical grid in its transition towards the new Smart Grid.

Deploying the Smart Grid until 2020 will require a tremendous amount of work and close collaboration from two domains, mainly Electrical Engineering domain (EE) and Information and Communication Technology (ICT) domain (see Figure 3.1). For a faster and reliable transition from the traditional electrical grid to the new Smart Grid, professionals with backgrounds in both fields are required [21]. The same collaboration is required in the research field, where joint research projects must be conducted in order to even the path between these domains. An important part of this research must focus on the security of the Smart Grid, and this must be done by treating security aspects originating from each of the base domains, together with new aspects that may appear. Communication and especially communication security is another important part of this research, the very large number of interconnected devices in the Smart Grid require fast, secure and reliable methods to exchange information between them and with the central management systems.

Many of the previous studies have either focused solely on computer communication and security issues, or on problems related to the electrical power engineering domain,

[1] http://ec.europa.eu/clima/policies/brief/eu/index_en.htm

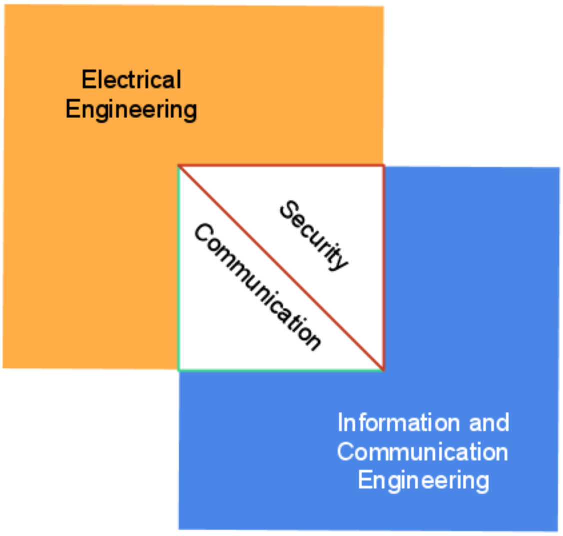

Figure 3.1: Cross domain research with focus on security and communication

often with few references to the other domain. In this chapter we make a survey of the literature relevant to each of the topics, with highlights on articles and studies which are important for our research.

3.1 SCADA security

As mentioned before in Section 2.4.3 the supervisory control and data acquisition (SCADA) system is the main component responsible for managing and controlling the electrical network. There is a lot of significant research to increase the security of these central management systems (SCADA systems) [22][23][24], as these systems tend to become connected to the Internet and also govern power production, meaning that any attack here may have serious repercussions.

Several research groups have also investigated *the state estimators* [15][25]. The state estimators are an important component in the process of creating the functional models of the electrical network by the central managements systems, and they can be used as a stepping stone for false data injection. For example, Liu et al.[15] present a type of attack targeting the sensors responsible of providing data for the state estimators in the SCADA system, and describe methods in which the measured data can be modified without triggering an alarm in the central management system. Although it is mentioned that the attack poses great difficulties for the attacker, as he needs to access directly the sensors which sometimes reside in remote locations (tall high-voltage poles in the middle of the field) , it is not impossible to be fulfilled. Performing such an attack may become easier in the future if these sensors will be connected to a faster communication network (such as the Internet) as proposed in [26].

Also, within the literature, there are some proposals for protection mechanisms and security recommendation regarding the SCADA system, as this becomes more and more interconnected with other systems, and the flow of information from one to another must be kept under strict control. Information security in SCADA systems is well documented by Göran Ericcson in [27][28][29], where the importance of security of all the components responsible of controlling and managing the electrical network is emphasized.

Another threat concerning SCADA systems is made by computer worms tailored for specific architectures and software versions [30]. The most important example is the recent Stuxnet [31] worm which targeted and infected a specific SCADA architecture by using zero-day operating system exploits, techniques to elude the active anti-virus systems, PLC (Programmable Logic Controllers) rootkit, peer-to-peer updates and complex process injection code. All these show that the creators of Stuxnet had very good knowledge on the internals of the targeted SCADA system, and considering a system to be secure just because it uses proprietary components is dangerous.

But these systems are not vulnerable just to external attackers, but the attacks can originate from internal sources, such as actual or former employees. In one cyber attack performed in Queensland, Australia in 2000 [32], a former disgruntled employee, still having remote access to the SCADA system (his remote access credentials were not disabled when he was fired), released 600,000 litres of sewage into public and private proprieties causing significant damage. Other examples of such cyber attacks scenarios concerning SCADA systems in water and electrical plants can be found in [33].

3.2 Electrical grid

When referring to the electrical grid there is extensive work focused on the stability of the grid, especially regarding frequency and voltage regulation in traditional electric network [16][17].

There are extensive studies regarding anomaly detection in the electricity transmission infrastructure. The purpose of these studies is to create algorithms and software tools which can learn the normal state of the network by observing the different electrical properties and then detecting deviations from this normal state by comparing the current running state with the learned one. For example in [34] and [35], Bigham et al. propose an anomaly detection framework based on invariant detection (such as range checkers, linear checkers and bus-zero-sum checkers) and heuristic approaches (artificial ants clustering) and the different components of the framework are tested in the case of data corruption and data loss. There is also a proposal for a framework taking in consideration an almost real-time reporting of data from the high-voltage transformation stations [26] which is intended to improve the process of state estimation presented in Section 3.1.

For a better control of the electrical network, the data provided by the new devices installed in the Smart Grid (e.g. smart meters) must be processed and integrated with the data from the SCADA system. This data can be used for example to better evaluate the stability of the grid and to keep the grid stable, the electrical company can use a technique called demand side management but which also known as load-shedding [36]. In exchange for financial benefits, such as cheaper energy, the customers agree to let the electricity distribution company control some of their large house-hold energy consumers (e.g. washing and drying machines, air-conditioning machines, pool pumps) operating parameters. The distribution company can turn off some of these appliances during high-load periods, thus reducing the stress on the electrical grid. In [37], Samarakoon and Ekanayake present a study in which these domestic appliances are controlled by a smart meter and they are used for primary side frequency response control, needed to keep the frequency in range, as presented in Section 2.4.4. Primary side frequency response control is a response mechanism which triggers in the first 30 seconds when the frequency of the grid starts to go awry. In the classical network, the power generators were responsible for the frequency response control but in the future it is possible that components in the distribution part of the network can participate in this task by becoming self-aware of the conditions in the grid, a condition proposed also by the Dynamic Demand project [38].

3.3 Advanced Metering Infrastructure

As the Advanced Metering Infrastructure (AMI) can be modelled as a large interconnected network (similar to the Internet), some studies [39][40][41][42] are covering the model and the functionalities of an Intrusion Detection System which can be used in AMI. The basic idea for the Intrusion Detection System originate from the ICT domain, but it needs to be tailored, by removing some obsolete functionalities and adding new ones suitable for this new type of network.

Recommendations regarding the cybersecurity architecture of AMI can be found in [43][44][45], and many of the recommendations originate from the ICT domain and refer to concepts such as Public Key Infrastructure and Trusted Based Computing. Regarding communication, there are proposals and studies covering encryption schemes used for communication or data aggregation in [46] and [47]. There is a strong demand for creating a Smart Grid which will provide secure and reliable communications from the early beginning and Hayden gives in [48] recommendations about creating a layered communication security model suitable for the new grid.

The most important devices in the AMI are the Smart Meters and in [9], Keeming and Roos perform a theoretical and practical analysis regarding these devices installed in Holland. Their conclusion is that even the security aspects are covered in the specifications, sometimes there is a deficient implementation of security in the actual physical devices.

A security problem regarding a type of communication modules used in the AMI devices is presented in [5]. Even though the communication performed using these types of modules is encrypted, extracting the encryption key from such a module is demonstrated. Using the same encryption key on a large number of devices reduces the security expenses of the company that installs and manages the meters, but also raises a security problem, because when a potential attacker discovers this single key he could take over a large number of Smart Meters.

In [49], McLaughlin et al. present a solution which may stop a large-scale attack to compromise a large number of smart meters. It is a software solution and involves encrypting the functions' return addresses when they are pushed in the stack at function calls. The solution is very ingenious because an attacker must create a specific virus or a trojan for every Smart Meter he wants to compromise.

Data privacy in AMI [50][51][52] is another important subject covered in the literature. With the introduction of new metering and monitoring devices new information about the behavior of the customers will be available for the electricity distribution company. These data can be used for good purposes, for example regulation of the power production [36], it can be used for marketing purposes or, if this information falls into malicious hands, someone can find out when a residence has inhabitants on its premises.

4

Investigation of the effect of possible attack cases involving smart meters

The overview given in Chapter 2 has its basis in smart meter security – we could see various alternatives implemented or proposed for creating a closed environment in the smart meter itself as well as encrypted communication infrastructure ensuring protection to sensitive information – and the tolerance levels of electrical energy quality, especially voltage and frequency, delivered to the the end customers. Let us now consider two scenarios that try to take advantage of the quality bounded electrical energy and the security implementation of the smart meters. Both scenarios are similar, in that the adversary takes control of a number of smart meters, but they differ in scale and the underlying property of the electricity network that the adversary will target. In the first scenario, the goal is to drive the frequency out of bounds – an attack that would require the adversary to control a significant amount of energy through the smart meters, causing imbalance at the level of the energy injection points that can propagate. In the second scenario, the goal is to vary the voltage in a small neighborhood – a simpler attack where the consequences would be localized to a chosen target. The second scenario is proven using a real-world simulation setup, and we present in further detail the initial conditions in Section 4.4.

Both of these scenarios require an interdisciplinary approach for the analysis of possible mitigation techniques and an understanding of their respective cost and weaknesses. For example, in the second scenario, from an ICT perspective, the smart meter can be installed with communication security already implemented with a higher cost of deployment, and from the Electrical Engineering perspective, an electrical engineer can install voltage regulators to specific weak points in the grid in order to mitigate the consequences of a possible attack. The aim is to combine those two approaches into building an even more secure network.

The focus is on the main steps of the scenarios and the impact of the attack on the grid, and not particularly on the details of the actual attack against the smart meters themselves. For that reason, before we present the two scenarios, we describe known weaknesses of smart meters that have been documented in other research.

4.1 Prerequisite: Taking control of the smart meter

The smart meter is a small embedded system, with a modular structure, with three complex components: the electrical meter, the processing unit and the communication module, and it has quite a number of vulnerabilities as described in Chapters 2 and 3. For example, some smart meters actually include a web server for query purposes that may open the path to direct attacks from the IT side. Given the number of exploits targeting web servers in the traditional IT setting, it is expected that the ones in the smart meter could also be exploited. In [42], Carpenter et al. describe a methodology to extract and reverse engineer the firmware from a smart meter to obtain valuable information about its internals, such as access passwords and communication encryption keys. The communication between the smart meters and the Central System (Metering Data Management System) is made through channels which have been proved to be prone to security breaches and a way to obtain the encryption key is available by exploiting implementation flaws [5].

Following this vulnerability in [5], another realistic scenario is described in [33], where by failure to obtain the wireless encryption key, due to system upgrade, the adversaries connected to a criminal network abuse the naivety of home users by means of social engineering and manage to get physical access to the targeted device, the smart meter, and to exploit the cascading effect of peer upgrades into controlling and causing havoc with the neighborhood's electrical energy.

Security flaws have also been discovered in the current implementation of smart meters and in [6] , McLaughlin et al. discuss tampering with the measurement device and problems related to the communication module with interception and injection of false messages. They also present a scenario where the injection of false malicious data lets the adversary gain different benefits from the system.

In order to gain access to a large number of smart meters, even if a remote attack is not possible, the attacker can employ social engineering. He can advertise a product or a "jailbroken" firmware, which supposedly will reduce the electricity consumption by a certain amount.[1] The attacker will gain twice from this: the revenues generated by selling the cost-reduction device and the opportunity to gain access to a large number of smart meters.

As can be seen, there are already several methods documented in literature on how an adversary can control smart meters. Given that this is a relatively new area of research, it is expected there are more vulnerabilities that are not known at this point.

[1] See for example the following URL for an instruction video how to do this with the traditional meter: `http://www.metacafe.com/watch/4659119/electricmeterhackhowtocutyourelectricitybillinhalf/`

4.2 Scenario 1: Frequency variation

Transmitting power from the energy producing facilities to the end-users requires transforming electrical energy into a form that has reduced cost and loss. One of the most used method is transforming into alternating current. In alternating current the direction in which the electrical charge flows is changed many times in the period of one second, usually following a sinusoidal pattern. For the European countries the frequency of the alternating current is 50 Hz, for USA is 60 Hz and some countries like Japan use both frequencies in different parts of the electric grid.

In the first scenario, the adversary targets the alternating current's *frequency.*A stable frequency is required for the stability of the electrical grid, and the whole electrical grid must be synchronized to the same frequency. However, the frequency in the electrical grid is closely dependent on the instantaneous energy generation and consumption, which thus must be balanced. If the frequency goes outside the $48 - 52$Hz range, total blackout may occur [38].

The Dynamic Demand Organization from UK [38] proposes a new technology called "Dynamic Demand Control" and its purpose is to help stabilize the UK national power grid by making home-appliances self-aware of the current state of the grid, mainly the power generation and power consumption. This is done by real-time measurements of the frequency in the electrical socket, and depending on the readings, the home-appliances should turn themselves on or off in order to maintain a frequency close to the 50 Hz value. If the electricity consumption is higher than the production, the frequency drops and if the production is higher than the consumption, the frequency rises. The home appliances would act like a big reacting back-up system which will have an important role in the stability of the grid.

As was mentioned before, one of the functionalities implemented in the current generation of smart meters is the remote ON/OFF switch. Its purpose is to facilitate remote energy turn off at a site (domestic consumer) without needing to send a maintenance crew to that location. In [53] Anderson and Fuloria raise and analyze the problem of improper use of the remote off switch. The ability to remotely turn electricity on or off for many customers simultaneously is new. Unfortunately, any capability can either be used as planned or misused by an adversary. Anderson and Fuloria point out that any vulnerability may lead terrorist organizations, environmental organizations and even individual criminals to be able to control this "feature," a feat possibly much simpler than to attack and destroy a power generation facility (the "traditional" way to cause a large blackout).

Controlling enough power from a number of customers can cause severe havoc in society, but the question is whether the attacker can force a larger blackout by having fine-grained control over the smart meters. The scenario setting is thus the following: an attacker takes control of a number of smart meters in a *large geographical area* by taking advantage of a remote communication weakness or through social engineering methods. By issuing synchronized commands to all smart meters to turn off their load, the result is an electrical network with excess generation, and no consumers. The central operators would at this point try to mitigate by reducing the power injection levels, but the attacker would in turn send a *turn on* command to all controlled smart meters, thus putting all the households suddenly back online. By such a technique, the attacker would try to destabilize the electrical network which could lead to a complete blackout. The effect of a successful attack of this type and magnitude is considerable, because re-establishing the functionality of the electrical grid could take from several hours to a few days. During that period there will be large areas without electrical energy, lack of communications, lack of heating in the winter, significant economic losses which will cause distress among the population.

The question is then how likely the worst outcome would be. Frequency variation can be observed in classical electric networks, based on the behavior of the individual consumers and their utilization pattern.[2] To prevent frequency variation caused by a quick demand of electrical power in the grid, the electrical network has reserves that can be injected. However, due to the significant cost of keeping these reserves in stand-by mode, the quantities available are usually estimated at the values required by specific standards/requirements. The success of the attack (in causing total blackout) is conditioned by its scale, i.e. the number of smart meters controlled (translated in volume of energy consumption) versus the number of energy producers (translated in volume of energy produced). Attempting this type of attack during the night would most likely not succeed – not much energy is consumed during the night and thus the attacker would not be able to control a critical mass – but the attack should be performed at times when there are already other stresses on the electrical grid (very warm summer day with air condition, or a very cold winter day for heating). The threat also depends on the structure of the grid in the country, and where power can be injected. A country such as Sweden, where a large part of the energy is produced by hydro and gas turbines is more resistant as these generation facilities have better response time compared to coal and oil electrical generators [54].

Even though the individual steps are already known, the above compiled and complete scenario with its possible consequences should be useful to both the researcher in security and the electrical engineers designing the networks. The large scale deployment of smart meters in the electric grid with the remote ON/OFF feature might open the ways to new types of attacks, if certain precautions are not taken. Research efforts into securing the smart grid is somewhat focused on the SCADA systems and the transmission network (see Chapter 3), but even attacks originating from the distribution side can have significant effects.

Another scenario which is restricted to a smaller area from the electrical distribution grid is described in the following section, followed by its simulation study.

[2] Demand surges can be quite significant when many customers act simultaneously, such as reported for the British royal wedding `http://www.guardian.co.uk/media/2011/apr/29/power-surge-royal-wedding-ratings`

4.3 Scenario 2: Voltage variation

While the target of the first scenario is a large part of the grid, the second scenario focuses on a small neighborhood, a *power island*, with only one power transformer. Following the definition of the power island, where this part of the grid in particular can produce electrical energy to become self-sufficient, some houses in the neighborhood have renewable energy production facilities, i.e. solar panels, installed on their premises. The renewable energy facilities produce energy for domestic consumption, but there can be periods when the production exceeds the demand and thus the excess energy is injected back into the network. Every customer has a smart meter installed, that in turn communicates with the data concentrator attached to the neighborhood power transformer, e.g. through a ZigBee network. The power consumption is reported to the data concentrator once every 15 minutes, and the smart meters can receive commands to turn on or off the electricity for the customer at any moment.[3] In this scenario all communication is encrypted with a symmetric encryption key but no other security mechanism is running on the smart meters. A graphical overview of the neighborhood is shown in Figure 4.2.

A common misconception about security is that *if* encryption is used, the network and the devices are safe from attacks. However, the devices may still be vulnerable to an exploit (buffer overflow), the protocols may not be well implemented, an oversight may lead to no change of the default settings, or there might even be an inside leak from the electricity company in question. As a parallel example, Stuxnet used valid signatures in its infection. [31]

Coming back to the attack scenario, the adversary is presumed to have a good knowledge of the smart meter deployment and the communication protocol used in the smart grid. Also in one way or another he discovered the shared encryption key used in smart meter's communication. Obtaining this encryption key allows the adversary to perform data sniffing and also data injection in the communication of the smart meters. He could extract useful information such as the energy consumption behavior of the residents in the neighborhood, peak energy consumption periods or the production capabilities of the renewable energy sources owned by the residents. Is his purpose to create havoc, by arbitrarily or systematically turning off and then on the electricity in the neighborhood. The discomfort caused by a power outage in a neighborhood is obvious : impossibility to use the electric house appliances during the outage, lack of electric light during evenings or early mornings, lack of external communications - if the land line phone is socket dependant -, temporary outage of the heating or air-conditioning system, etc. The electrical effect of such an attack is an interesting fact to be known. If an under voltage will cause minimal damage, because, in the worst case, some appliances will fail to work correctly, an over voltage can cause serious damage to the appliances, even make them unusable afterwards if not properly protected.

The purpose of this scenario, further explored in the following sections, is to deter-

[3] The smart meter has its own power supply, so it does not depend on whether the electricity in the house is on or off.

mine if an attacker can make the voltage of the network go outside the tolerance limits of +10% and -15% of the nominal value, which for most European countries is 230V, and what are the energy requirements to do so.

4.4 Simulation environment

To demonstrate the feasibility of the voltage variation attack scenario, we use the evaluation version of the commercial software tool named PowerWorld Simulator [55].

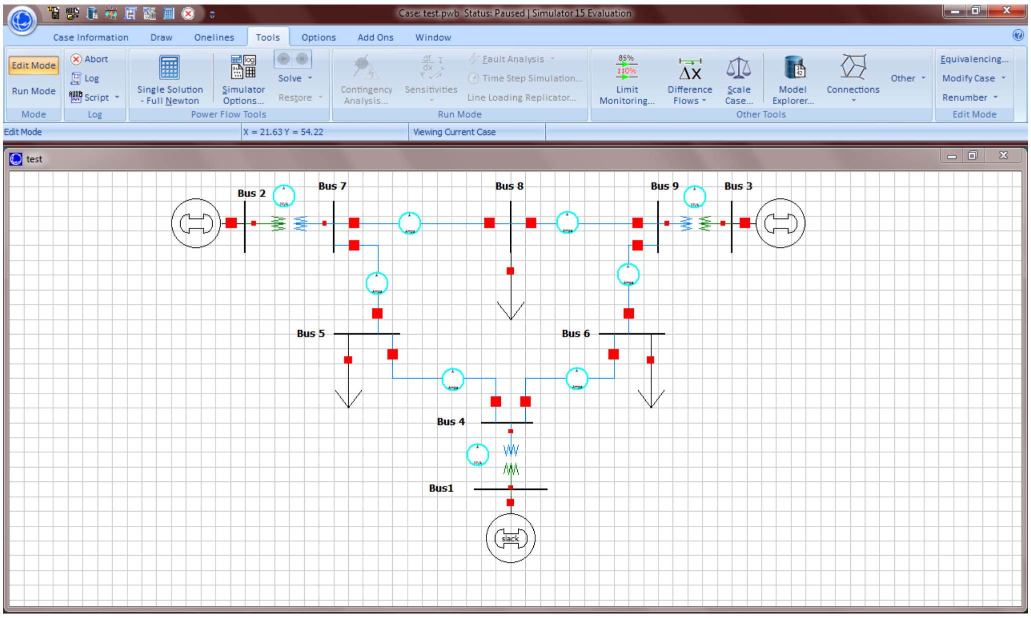

Figure 4.1: PowerWorld Simulator graphical user interface.

This software suite provides the necessary means to model the realistic grid configuration of the power island described in the previous section. Modelling transmission lines, generators (power sources) and loads is possible by mentioning the right specifications in the tool, that provides modules able to solve the resulting power flow equations of each node and display the end result - voltage and power levels flowing in and out of each node.

The upper part of Figure 4.2 presents an intuitive overview of the neighborhood while the lower subfigure shows the overview of the resulting electrical network model used in our simulation. As it is specified in the legend, the thick vertical lines are called buses and each bus belongs to a residence. The renewable energy sources (solar panels) are modelled as generators while the consumers in the residence are modelled as loads. Bus #1 represents the electrical distribution station which serves the neighborhood.

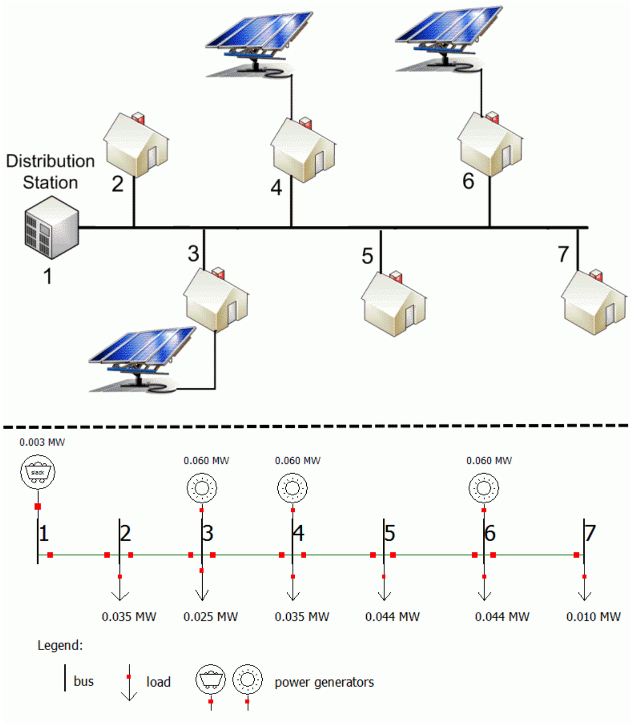

Figure 4.2: Neighborhood overview (top) and electrical network model overview (bottom)

4.5 Setup and simulation of voltage variation scenario

The neighborhood is a typical country-side distribution grid, with several buildings and their facilities served by one power substation (marked as node one in the figure). The six buildings (numbered two to seven) have a relatively higher than normal individual instantaneous energy consumption ranging from 2 to 60kW. There are three renewable energy production facilities in the neighborhood, connected at chosen nodes: three, four, and six respectively. These facilities can produce more energy than required for the local neighborhood, so the surplus is injected into the electrical network. The injection of energy produced - better said transformed according to the process of obtaining electrical energy - by the green sources is not done directly, but through energy storage equipments that are used to store the energy produced during the day, the same quantity of energy is evenly extracted from these during the 24 hour period. The bars numbered from one to seven in the lower subfigure are called *buses*. Buses are points in the electrical system where certain electrical attributes such as voltage, power and current can be evaluated ("p.u." signifies the voltage per unit value of each bus). Every building that consumes energy is modeled as a load, every renewable energy facility is modeled as a generator and the electrical lines connecting the nodes are modeled as transmission lines with proper loss (see Apendix A). We utilize four real-time daily consumption profiles for the customers adapted from [18]. In Figure 4.3, the consumption profiles are based

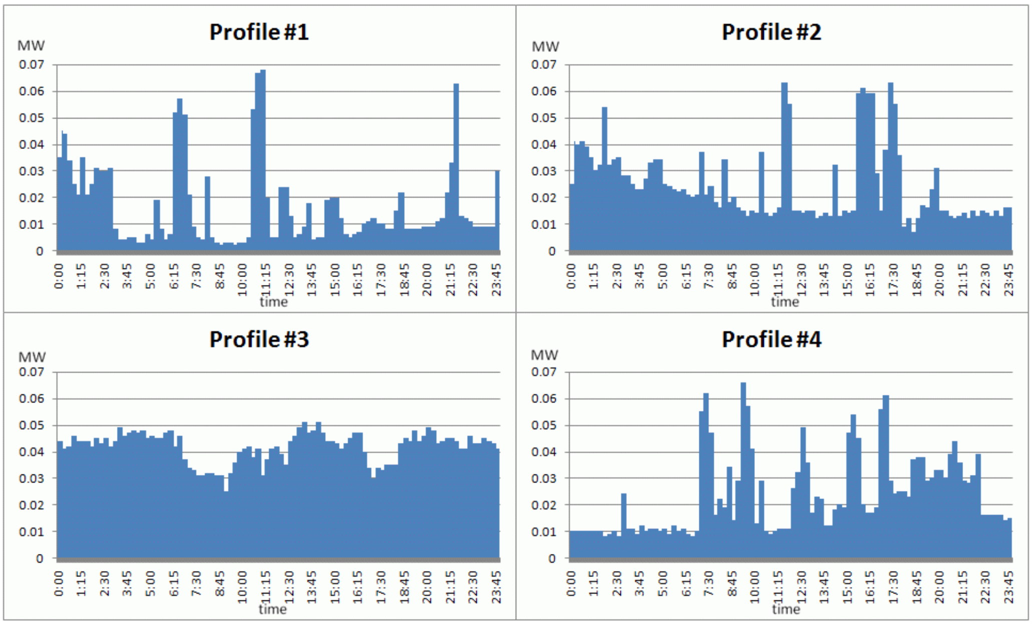

Figure 4.3: The consumption profiles for four different customers

on 24 hour consumption patterns with 15-minute interval measurements. These profiles are characterized by peaks during the rush hours (in the morning, at lunch and in the evening), depending on each household's appliances in use. For example profiles #1 and #4 belong to customers who consume more during morning, lunch time and during dinner (when the kitchen appliances such as oven or kettle are running). Profile #3 belongs to a customer who spends most of his/hers time at home and during this time he/she is an intensive consumer of energy. Profile #2 belongs to a customer that is more active during the night, energy consumption-wise speaking. It may be the case that this is the period when the electrical energy is cheaper, so charging the electrical vehicle or using home appliances such as washing or drying machine seem more price reasonable. We choose customers #2 and #4 to use the consumption profile one, customer #3 use profile two, customer #5 and #6 use profile three and finally customer #7 use profile four (chosen arbitrarily).

Attributing a consumption profile to each load has been done arbitrarily: Load #2 and #4 have the consumption of Profile #1, Load #3 of Profile #2, Load #5 and Load #6 of Profile #3 and Load #7 of Profile #4. Implementing those profiles into the loads of the simulation has been done by increasing or decreasing the power consumption of the load on a 15-minutes basis matching the consumption profiles described earlier.

In real world conditions, distributing electricity to end consumers is done either traditionally by using overhead lines, or by underground lines that offer more security. We consider the two cases for the simulation, one in which the distribution lines are placed overhead and one in which the lines are placed underground. As mentioned in Section 2.4.2 the type of line placement causes a different voltage drop on the line and thus the transformer needs to adjust the power injected in the neighborhood to maintain nominal voltage. We expect that the effects of the attack will be more significant in the overhead line case. With this in mind, deploying an electrical grid is done with power quality design considerations, i.e. voltage magnitude should respect the power quality standards [16]. In the following, results obtained from the simulation steps are presented, with the notification the any modification to the load profiles or grid configuration will lead to totally different effects. The simulation is configured to run during 24 simulation seconds, equivalent to a 24h normal day.

4.5.1 Normal running conditions

In Figure 4.4 we present the voltage magnitude variation of Bus #7, i.e. the load for customer seven, for the case of underground distribution line. The grid's behavior (loss, power injections, loads) is responsible for the voltage maximum and minimum points seen in the figure. The voltage peaks are the result of large amounts of power in the grid and few consumers, while the voltage minimum points are a result of many consumers and little available power. In normal running conditions, the voltage profile of Bus #7 respects regulations and the voltage magnitude never goes beyond +8%/−6% of the nominal value.

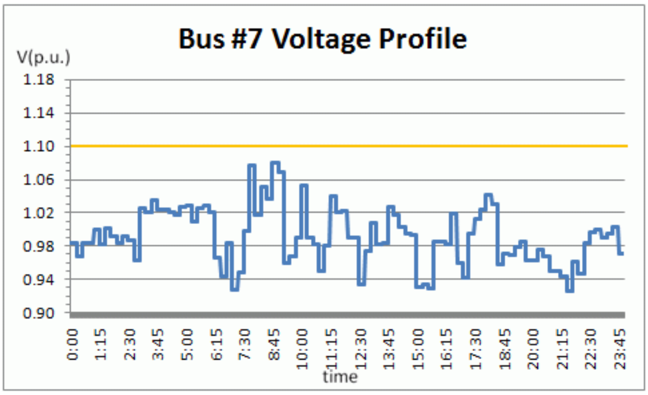

Figure 4.4: Voltage level at Bus #7 during the simulation for the underground distribution line

Figure Figure 4.5 represents the voltage magnitude variation of Bus #7 also, but this time, in the case of the overhead distribution lines. The boundaries are kept for the voltage magnitude in this case, but it can be observed that the voltage magnitude is slightly higher than in the previous case, which is caused by the overhead line properties. The simulation respects the initial conditions of the underground distribution line test case,by keeping the same variation in the loads and the same values in the power generation facilities. This allows a good comparison between the underground and overhead power distribution.

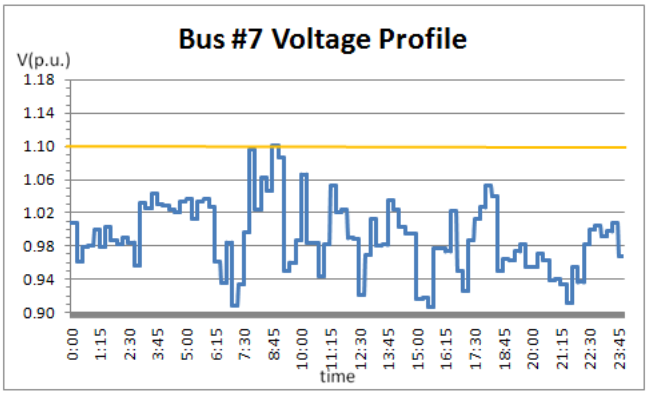

Figure 4.5: Voltage level at Bus #7 during the simulation for the overhead distribution line

4.5.2 Modelling the adversary

The goal of the adversary is to vary the voltage magnitude of Bus #7 outside the safety zone of ±10%. This is achieved by careful manipulation of the smart meters in the neighborhood to shut down power systematically and observing the effects on the most attractive target. Prior to the attack itself, the adversary needs to know in advance the consumption and voltage profile of each customer in order to build an overall idea of when situation close to the limit occur in the neighborhood. This could be possible if the network allows sniffing messages sent by the smart meters, or by directly reading each smart meter's memory using a malicious piece of software previously spread in the neighborhood. The simulation will be conducted again for both distribution lines type e.g. underground lines and overhead lines.

Underground distribution line case

In a first attempt, the attacker gains control of the meter controlling the load on Bus #5. The attack is then launched at an appropriate point in time, for example at the voltage peak observed between time $7 - 10$ when the grid is vulnerable; there is then a high demand for energy (people are preparing for going to work) and the generators must compensate and push more power into the grid. If the attack is timed correctly, Load #5 will be interrupted during the established period, leading to more power being routed to the other buses. The result is shown in Figure 4.6, where the highlighted area represents the time of the attack.

The voltage magnitude barely goes up 2% of the established barrier at 1.1 volts per unit and only for very short period of times. This may not be enough to cause damage to the target and the attack cannot be deemed successful.

In a second attempt, the attacker gains control of an additional smart meter (the one controlling Bus #6). Following the same attack procedure as described above for Bus #5 , both smart meters are used to abuse their control capabilities to turn off the

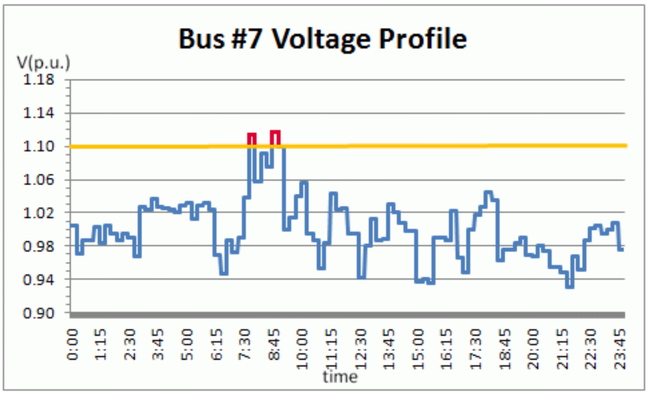

Figure 4.6: Bus #7 voltage after launching the attack on Bus #5 with the attack period emphasized with the square box. The safe voltage limit is at 1.1 V(p.u.) and the voltage is normalized. The underground distribution line case.

power to the respective customers. The result can be observed in Figure 4.7.

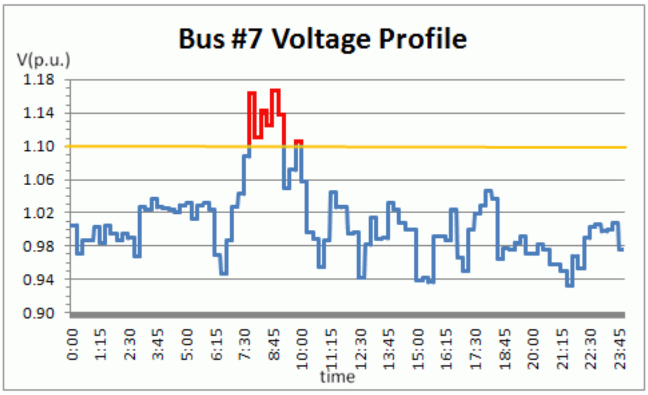

Figure 4.7: Bus #7 voltage after launching the attack on Bus #5 and #6 with the attack period emphasized with the square box. The safe voltage limit is at 1.1 V(p.u.) and the voltage is normalized. The underground distribution line case.

This time, the adversary manages to drive the voltage magnitude to peaks of +17% with a constant average value above the normal +10% between time 7 and 10. This increase in voltage should be enough to cause major damage, both physical and economical, to the customer in the absence of voltage regulators.

Overhead distribution line case

As described in Section 4.5.1, the second power distribution case is using the traditional overhead lines. Because of the different line properties, results are expected to be different as well, but with stronger effects. The scenario follows the same steps as described for the power distribution in underground lines and the goal remains unchanged.

After gaining control of the smart meter at residence #5 the voltage magnitude profile from Bus #7 modifies as showed in Figure 4.8. Adding the smart meter from residence #6 will cause the voltage at Bus #7 to go as high as 22% of the nominal voltage. This can be observed in Figure 4.9.

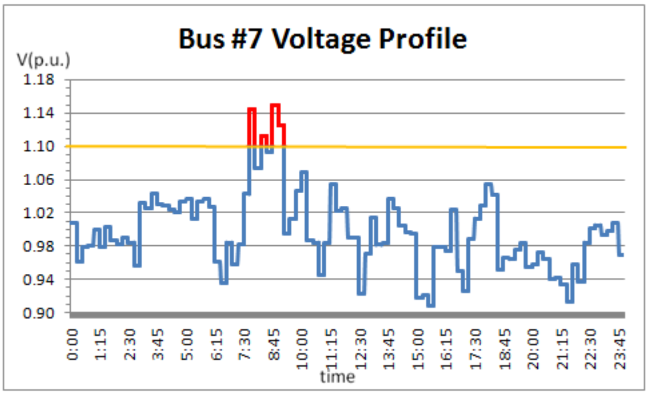

Figure 4.8: Bus #7 voltage after launching the attack on Bus #5 with the attack period emphasized with the square box. The safe voltage limit is at 1.1 V(p.u.) and the voltage is normalized. The overhead distribution line case.

In both cases the effects of the attack are much stronger for power distribution in overhead lines as opposed to underground lines. By comparing Figures 4.8 and 4.6 of the first attempt where only Bus#5 is taken over and Figures 4.9 and 4.7 the voltage magnitude tops values nearly 5% higher.

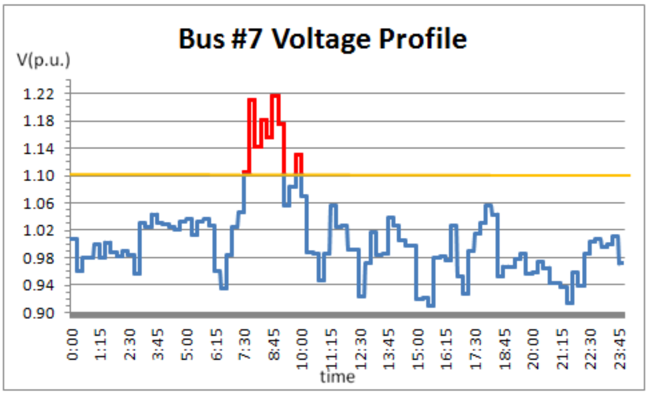

Figure 4.9: Bus #7 voltage after launching the attack on Bus #5 and #6 with the attack period emphasized with the square box. The safe voltage limit is at 1.1 V(p.u.) and the voltage is normalized. The overhead distribution line case.

This environment is more susceptible to attacks and much more attractive form the perspective of an adversary, since the voltage magnitude can be driven outside the standardized boundaries with less power controlled and thus less smart meters controlled.

4.6 Conclusions from the simulation

As an observation, the number of smart meters that need to be compromised is not in direct relation to the total number of smart meters in the neighborhood, but with the electrical power behind each smart meter. For example, a smart meter that controls a high-load household is more attractive for takeover since the strain it can reflect into the electrical grid is higher. The demonstration of the threat is clear in that voltage magnitude tolerance levels have been exceeded to the targeted household, managing to at least make electronic devices inert. The side effects of this should not be neglected as there will be other households in *darkness* during the attack without the central operator knowing it and hard to check or even pin-point the problem. There is also the aspect of economical loss that the electrical distribution company or the owners of renewable energy facilities suffered when there was a lot of power injected into the grid and no one to consume it.

Although the prerequisites of the threat scenario seem not too complicated, such a scenario can be easily avoided. The first step in the preventing process is a careful deployment and control of the ON/OFF feature. A large scale ON/OFF command should not be issued by only one source, but it should require the consent of several parties and a strong protocol should be developed for issuing a large scale ON/OFF command. This feature will complement well the already existing security features on the current generation of smart meters, and can be the scope of future research, as it could be sketched as a distributed algorithm.

5

Using AMI data to detect and mitigate attacks

The introduction of new metering equipment in the electrical distribution infrastructure and the connection of this equipment via a communication network provide more accurate and frequent data readings. Before this, when the electrical meters were analogue, the data were typically read once a month by an operator employed by the electricity distribution company. The operator would need to go from residence to residence and read the consumption index from the meter's display, or if the meter was digital, the operator would use a specialized tool to access the data. The data were carried back to the data gathering center, usually situated in one of the electricity distribution company's buildings where they would be added to a database. This database would then be used by the billing department, the maintenance department, the marketing department or any other department that would need the data. As mentioned before, the frequency of data reading was monthly, even for customers who would have a different tariff based on day or night consumption. In this case the operator would read two indexes.

The new metering equipment can provide data with a higher granularity, at a rate of one reading every 15 minutes. With the introduction of new communication algorithms and the extension of the communication networks [56] [57], the goal is to achieve real-time readings in the future. Such data have a higher degree of complexity because, apart from the classic consumption index, they also contain information about power quality, power outages, or possible the renewable energy sources owned by the customer. This additional information in turn paves the way for new usage possibilities, for example in power production management, improving AMI security or for marketing purposes, as described below.

5.1 Analysis of gathered data

As mentioned before, AMI allows for the possibility of gathering data from points in the network where, before this, the electricity distribution company was blind. There are many actors on the scene that find these data useful. The first one to use these data is the electricity distribution company. The data is currently primarily used for billing purposes, with the mention that billing can be done now by using time-based tariffs, so the customer is more interested in using his appliances during the time periods when electricity is cheap. The data can be used for a better management of the electricity production [56], because better knowledge of the way the electricity is consumed and about the renewable energy sources owned by the customers will improve the energy produced/energy consumed ratio by minimising the loss of electrical energy in the network. The problem with all these data is that they are very sensitive, and a lot of information about a customer's behavior could be inferred [51]. If we assume the honest use of the data on behalf of the electricity distribution company, we cannot assume the same when thinking that these data can fall into malicious hands.

The second actor that can make use of these data is an attacker who can have various objectives. In [6] McLaughlin et al. define four types of attackers that would be interested in stealing energy from the Advanced Metering Infrastructure. Knowing the behavior of dwellers in a residence can be a very useful information for a criminal group, because a robbery can be conducted when there is no one at the residence or the dwellers are away on holidays. These data can be used to conduct an attack similar to the voltage variation attack presented in Chapter 4, as the attacker can create the power consumption profiles of the targets. These are just a few examples on how the data can be used maliciously, and there may be even more, only limited by the attackers' imagination.

A third group that can use these data is the research community. Many solutions for problems in the electrical network originated from the research community. For example, research involving the state estimators used for building the electrical network models in the SCADA system is covered in [15], [34] and [35]. In [56], Meliopoulos et al. propose the integration of data originating from the smart meters in state estimation system, similar to the one used in the transmission network. Developing and tuning such a system requires the utilization of real data, data which need to be provided by the electricity distribution company. Providing these data in raw format is unacceptable because data can be traced back to the customer, where, for example, the smart meter serial number can divulge where the smart meter is installed. The solution to this problem is to anonymize the data that is provided for research purposes.

5.2 The need for data anonymization

As mentioned in Chapter 3, an important subject in the Advanced Metering Infrastructure is data privacy [50][51][52] and a solution to maintain the privacy of data is to use data anonymization. In [50], Efthymiou and Kalogridis propose a framework in which the readings from the smart meters are performed via a 3rd party escrow mechanism which provides authenticated and anonymous data readings. By using this system, the data would be anonymous even for the electricity distribution company and only the data required for billing or marketing purposes could be associated with a specific client. Although this appears to be a good solution, implementing it may involve additional costs which would need to be supported by the electricity distribution company.

If data about consumption and power quality are needed in the research community for studying the power profiles or building specific tools, sensitive information such as smart meters serial number or client ID are not so important and are better to be kept confidential. These data can be released in an anonymized form. The simplest way to perform anonymization is achieved by using a random function, replacing every original serial number or customer ID with a random value. Before actually replacing the value, the random value must be checked against all the anonymized values generated before, in order to avoid collision. The method is simple, but it lacks consistency. For example, consider the case when additional data need to be added to the original set or the same mapping between the original and anonymized values is needed in two different sets.

Solving this problem requires the use of an one-way anonymization function which provides a one-to-one mapping and is consistent across different traces. One solution comes from the networking domain where in [7], Fan et al. propose an anonymization solution for IP addresses that can be used by network trace specialists. This solution is called Crypto-PAn and it stands for **C**ryptography-based **P**refix-preserving **AN**onymization. Crypto-PAn provides one-to-one, prefix-preserving, consistent across traces and cryptography-based anonymization for IP addresses. The consistency across different traces is achieved by using the same secret key in a cryptographic pseudorandom mapping function. They propose a theorem, called Theorem 1 which defines the anonymization function as:

Theorem 1. *Let f_i be a function from $\{0{,}1\}^i$ to $\{0{,}1\}$, for $i = 0{,}1{,}...{,}n-1$ and f_0 is a constant function. Let F be a function from $\{0{,}1\}^n$ to $\{0{,}1\}^n$ defined as follows.*
Given $a = a_1 a_2 ... a_n$ let: $F(a) = a_1^{'} a_2^{'} ... a_n^{'}$, where $a_i^{'} = a_1 \oplus f_{i-1}(a_1 a_2 ... a_{i-1})$ and $\oplus$ stands for the exclusive-or operation, for $i = 0{,}1{,}...{,}n-1$.

They claim that the function F is a prefix-preserving anonymization function and that a prefix-preserving anonymization function necessarily takes this form. The proofs for both of these statements are provided in [7] and are not included here.

To make this scheme cryptography-based they instantiate the f_{i-1} function from Theorem 1 with $L(R(P(a_1 a_2 ... a_{i-1}), k))$ where L returns the least significant bit (the last bit in a string of bits), R is a cryptographic pseudorandom function, P is a padding function that expands $a_1 a_2 ... a_i$ in order to match the block-size requirements of R and k is the cryptographic key used in the R function. Table 5.1 contains the logical table for the XOR operation, which is the bit-wise operation used between the result of the above function and the original value.

p	q	$p \oplus q$
0	0	0
0	1	1
1	0	1
1	1	0

Table 5.1: XOR (Exclusive OR) Logical table

5.3 Suggested anonymization solution for smart meter serial number or customer ID

5.3.1 Anonymization procedure

In this section we are going to present a solution for smart meter serial number or customer ID anonymization, a solution which is derived from Crypto-PAn [7].

We presume that the smart meter serial number is an decimal number (possibly pre-padded with zeros) with D decimal digits. For example such a serial number can have the form of 456789, 000012345678 or 999123456789. Our purpose is to anonymize this serial number by mapping it to another number, similar in format (the number of decimal digits). So the input will be a fixed-size string containing decimal digits. In order to obtain a similar string at the end of the anonymization procedure, we are going to interpret this string as an integer number which will be transformed to binary form. We define B as the number of bits required to represent the highest value of a D decimal integer. For example, the minimal number of bits required to write 999 is 10 (2^{10}=1024 and 2^9=512). So the first step in the anonymization procedure is to convert the original object into a binary unsigned integer and pre-pad it with zeros (if necessary) until it reaches the length of B. For the pseudorandom function R we choose to use the AES encryption method [58], based on Rijndael which is the pseudorandom function originally proposed in Crypto-PAn. The size of the key for AES is 128 bits while the size of the padding is also 128 bits. The key and the padding can be concatenated and distributed among different research groups as a 256-bit key.

The result of the anonymization procedure is a B bit unsigned integer which needs to be converted back into integer form and pre-padded with zeros (if needed) until it has D decimals.

We have made two implementations of this procedure, using two different programming languages, JAVA and Python, and the code for both of them can be found in Appendix B.

Pseudo-code 1 The pseudo-code for the anonymization procedure.

```
for i=1, i<len(p), i++:
    f[i] = p[ 0:i-1 ]+k[ 128:255-i+1 ] //128 bit padded block
    rs   = R( f[i] , k[0:127] ) //encrypted block using AES
    otp  = otp + L(rs) //appending the least significant bit
                      //of rs to otp
s = p xor otp //obtaining anonymized result s

where:

f(0)=0  -> a constant function, and otp is initialized with f(0)
otp     -> B bit one time pad, used for xor
p       -> B bit plaintext string to be anonymized
s       -> B bit anonymized result
k       -> 256 bit key (128 bit cryptographic key + 128 bit padding)
```

5.3.2 Preserving the prefix

Due to the integer to binary transformation, the prefix-preserving property is not satisfied if the transformation is made on all the B bits. One example is shown in Table 5.2. The property can be satisfied if the length of the prefix is known beforehand so the procedure can be applied to a string of bits of length B which is a sub-multiple of the length of the prefix. This implies that the structure of the smart meter serial number plays a significant role and must be accounted for in the anonymization process to preserve the prefixes. Prefix-preserving can be very helpful when analyzing clusters of smart meters from the same area for example, if the prefix of the serial number represents a specific area.

Original serial number	Anonymized serial number
000000245996362274	277241976953513947
000000224627243464	277241981321211102
000000057410347131	277241881814790385
000000198053243924	277241931107945192
000000674927344848	277242359738238053
000000317160084543	277242071338556968
000000570834029433	277242399684700719

Table 5.2: Results of non-prefix preserving procedure on a 18 decimal serial number

5.4 What can gathered data tell us?

An overview of the smart grid has been done in Chapter 2, where we describe a smart meter as a small embedded device capable of processing, storing and transmitting data to a Management Data Center through various communication media following special protocols. Currently, the main purpose of centralizing consumption information is to automatize the billing process, allowing customers to observe their consumption with much greater detail. Besides billing, consumption data can prove to be invaluable when careful analysis is conducted with regard to planning and efficient management of resource utilization or, from a security perspective, data can be used for building and implementing the detection rules for defense mechanisms.

Assuming that an energy distribution company keeps track of the smart meter readings for its entire customer base, a research study has to handle a huge amount of data. From this a variety of consumption profiles can be built to aid in the energy planning phase and in the same time to reduce the loss that results from the ratio between energy production and energy demand. Since collecting information is not always error free, we must assume a fair rate of anomalous reporting stemming from either an incorrectly installed meter equipment, device tampering, communication medium corruption or even an outside attack. Before forwarding the information to analysis for further study of the consumption profiles, a stage of classifying anomalies in the recordings must be conducted, identifying and assigning meaning to each abnormal entry in such way that discarding is not done on the grounds of data invalidity.

In the following sections we will present a description of these data, with the scope of classification, detection, and potential uses of the study of individual customer profiles; we will also analyse how building up clusters of customers can aid the development of mitigation techniques.

5.4.1 Data from one smart meter

When working with huge amounts of information, containing consumption characteristics (serial, time stamp, consumption index, etc.), the first step before concluding anything about the data is to look at each individual recording and decide whether it is valid. This calls for a sanitization process that recompiles the data into a form that allows lean graphs in consumption profiling. But, as mentioned before, these abnormal data may signal problems that stem from a wide range of causes, from software bugs and errors to communication paradigms, faulty equipment deployment or even attacks.

As such, following data rules, any *unrecognizable data values* or negative values may be the result of a software bug in the management center or may be due to the high variety of metering equipment installed. Each smart meter may be running a different firmware version and in the reporting data process, information is wrongly interpreted. Other reasons may be found on the metering equipment due to faulty wiring during installation, equipment defects or errors, but also an attack should not be excluded. Acknowledging these data may allow energy distribution companies to track down the faulty equipment for repair or replacement purposes - this may be the case of rogue, zombie or lost smart meters. However, having a wide range of factors that can cause an unrecognisable data value makes the identification of the precise reason difficult.

Focusing on the sanity of consumption indexes for each customer, we must verify that the *index value does not decrease over time.* The time issue in this case may stem from the type of communication infrastructure. If we deal with a distributed architecture, synchronization problems may affect message order and time stamping, thus making detection based on consumption indexes very difficult. In a centralized architecture, where each message is time stamped at the meter equipment, detection becomes trivial, but we cannot say the same thing about the cause, as it may have roots in software bugs or an attack attempt for energy theft.

Time gaps in recordings, assuming that the time collection primitives are known, can be easily detected by checking the time stamps. As always this problem can be caused by the communication infrastructure, where messages can be lost, but when the study is expanded to the entire records of a customer, this may show periods of energy shut down caused by overdue payments or blackouts. An adversary can also benefit from this, if the packet containing the record is redirected to his own server.

Depending on the time collection primitives, other interesting sets of data are *constant consumption indexes* over long periods of time. Analysing these data for individual customers, may show periods of leave or vacation, assuming that they have unplugged all of the electronic devices in the house, or that the metering equipment fails to update the index value. The adversary perspective says that it may be the case when energy theft is performed through specialized device tampering tools. A particular case of this is for a replay attack, when the adversary tries to discover the security features of the communication, and resends a message he intercepted to the data center.

With the help of records from each individual customer, a time consumption profile can be built (same as the profiles we described in Chapter 4), allowing a better description of their type of activity (active during night, worker during day, etc.). Peak and trough consumptions over long time periods can be compared to identify deviations from the normal behavior, indicating either a problem in the grid if the peak is significantly over the limit, or energy theft if the consumption profile average levels drops significantly. For further studies, developing mean consumption profiles from the individual customers helps in creating a global data repository for other research purposes. A variation for average values will be used in the next section to cluster customers in different categories.

5.4.2 Data from clusters of smart meters

The last section showed how data from one smart meter can be used to infer information about a specific customer. In this section we propose some ideas about grouping the smart meters into clusters and describe what information can be inferred from this. We can cluster the smart meters based on geographic location (smart meters from the same area, neighborhood, block of flats, etc.) or on similar energy consumption patterns.

Clustering smart meters depending on the geographic location can be done in different ways. The simplest way is to find the customer's address where the smart meter is installed, as this is recorded at the electricity distribution company. If the address is confidential, information about the location can be extracted from the smart meter's serial number or customer ID, if part of it (the prefix) is consistent for smart meters installed in the same area. If the serial number is anonymized, the area-based clustering can still be performed if a prefix-preserving anonymization procedure is used, such as the one presented in Section 5.3. If no information about the location can be extracted from the smart meter serial number (the serial number's prefix does not infer location or the anonymization is made by using a random anonymization function), similarities in the power consumption profiles (like power outages in the same time period) can be used to link smart meters to a specific area.

Clustering smart meters based on power consumption profiles can have several benefits. Grouping customers with similar consumption patterns can help in detecting problems in the electricity distribution network as presented in [59], and abnormality detection can be further used in the process of fraud detection [60], together with data mining [61].

One of the base clustering methods used is k-means clustering [62]. The idea in the k-means clustering is to partition a set of n observations into k different clusters, such as every observation would belong to the cluster with the nearest mean. The initial data are a set of observations $X = (x_1,x_2,...,x_n)$ where each observation is a d-dimensional real vector. This set must be partitioned into k clusters, such as k$\leq$n, so as to minimize the within-cluster sum of squares (WCSS):

$$\arg_S \min \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} ||x_j - \mu_i||^2,$$

where μ_i is the mean of the points in S_i. The mean of every cluster is called a *centroid.*

By using k-means clustering, smart meters can be grouped in tiers, e.g. low, medium, high, based on their mean energy consumption during a period of time (hour, day, month). The clusters obtained this way can be observed during several periods of time for changes in their structure, for example if a smart meter persistently belongs to a specific tier or it changes different clusters during these periods. This change can be caused by a variation in the customer behavior, a problem in the electricity network, a faulty meter, a fraud attempt, etc.

5.4.3 Scenario setting

Let us take for example a case where we have consumption data readings for a number of smart meters. The data were recorded with a frequency of one reading every 15 minutes, so for a 24-hour period we have approximately 96 readings for each individual smart meter. Every smart meter can be identified by an unique serial number. The format of the recording file can resemble the one shown in Table 5.3.

As mentioned before, analysis on these data can be conducted on individual smart meters or on clusters of smart meters.

Smart meter serial number	Timestamp	Index in kWh
...	...	...
000000245996362274	2010-08-08 13:44:06	455.65
000000224627243464	2010-08-08 14:03:34	675.34
000000057410347131	2010-08-08 14:14:21	347.22
...	...	...

Table 5.3: Recording file format

Analyzing individual smart meters

Using the information from the scenario setting described in the previous section together with the tools from Section 5.4.1, we can define a quick method to detect customers which are potentially performing energy theft.

The simplest way to detect these customers is to verify that the index values are increasing over time and to check also the frequency of index reporting. Any time gaps in index reporting can be caused by a equipment failure, or a deliberately obstruction of the communication process. When checking the index, one must also have in mind that an index which is decreasing during time can be caused by a customer who has a renewable energy source and its injecting the surplus back into the network. If this customer is consuming less than he is producing, the difference is sold back to the distribution company, and the quantity of energy would be registered as a negative value by the smart meter.

Expanding this detection technique could include two other verifications. Checking for periods when consumption is zero or when the consumption remains constant over long periods of time can indicate a faulty meter or an energy theft attempt. Also, a double recording can indicate a problem in the communication process or a replay by a device which mimics a smart meter.

Building individual consumption profiles for each customer helps to identify the peak and lows in consumption over time. The history of consumption helps both the customer and the electricity distribution company. The customers become aware of their peaks in consumption over time, and if this is coupled with a different time of use price for energy, it may help to even the consumption spikes, thus lowering the stress in the grid. Consumption profiles can also provide sensitive information, such as time intervals when the customer is away, equivalent to the period with a constant minimum consumption. Also the level of a constant peak load may reveal some private information about the customer, such as the social status, i.e. very high consumption levels may point to a very large set of electronic devices, thus indicating a wealthy customer.

Having a consistent history for each of the customers helps in building an accurate model for each of them. The longer the period of observation is, the more accurate the

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
0003	0001	0010	0006	0002
0030	0004	0017	0007	0005
0040	0009	0021	0008	0011
0050	0013	0032	0012	0018
0059	0014	0046	0015	0020
0060	0016	0048	0019	0029
0064	0022	0061	0023	0037
...	...	...	...	...

Table 5.4: Division of smart meters in different clusters

model may be. So, if a consumption peak is higher than the observed normal, an attack may be under way, or if a dimming of the average level is observed for longer periods of time, it may be the case of energy theft.

Analyzing clusters of smart meters

Lets presume that we have consumption data from 1,000 smart meters from the same area. Each smart meter corresponds to one residence in the area and each smart meter has an unique serial number from 0001 to 1000. We want to divide these smart meters into 5 different clusters, based on each meter's daily average energy consumption. In the absence of real data, we have generated consumption values for each of the smart meters. The graphical result of the k-means clustering algorithm is shown in Figure 5.1, while Table 5.4 shows the serial numbers of the smart meters in each cluster.

Repeating the k-means clustering algorithm over different days and observing deviations of the clustering process can be a good way to detect irregularities in the energy consumption process.

Applying the clustering algorithm for power consumption profiles can be used to detect frauds such as electricity thefts, like the method presented in [59]. Also, knowing the power consumption distribution is an important prerequisite in implementing load-shedding [36]. Shaving the energy consumption peaks through load-shedding, and distributing the consumption evenly over a 24-hour period in a power island, can be a mitigation technique for the voltage variation attack which is presented in Chapter 4. Having an almost constant consumption profile over a 24-hour period would hinder the attacker's job to find a suitable target and maybe it will make him search for one in another power island. This would also release the stress on the electrical distribution station which would not have to support high consumption peaks.

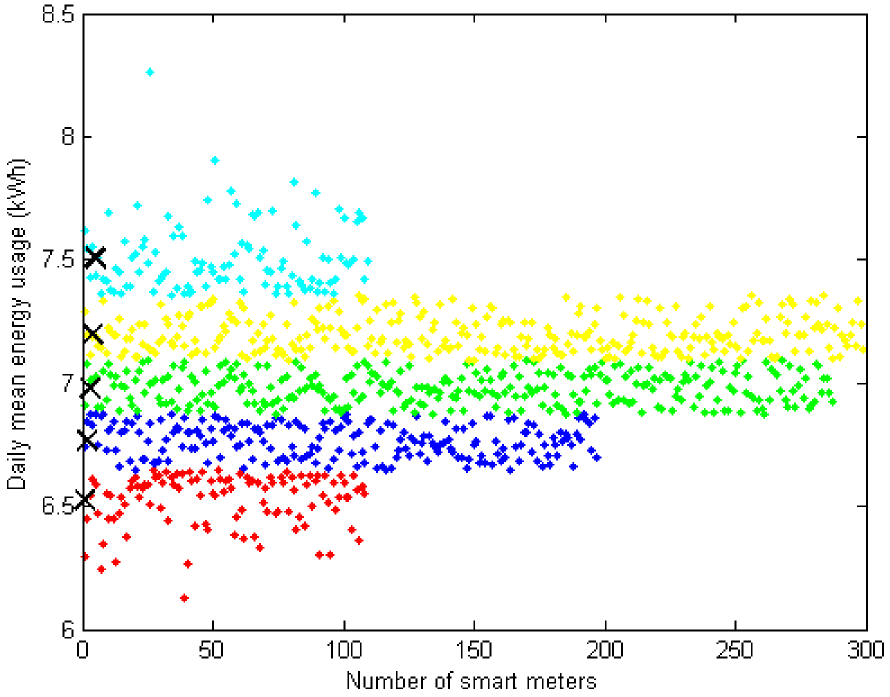

Figure 5.1: Dividing 1000 smart meters into 5 clusters, based on the daily mean energy consumption. Symbol × marks the centroid of every cluster.

5.4.4 Using data from distributed sources

Previous sections dealt with data stored in a central data base. This is a simple case and can be applied when the size of AMI is relatively small and data can be quickly transmitted and stored in a centralized manner. When the number of customers served in AMI grows, gathering, storing and processing data in a centralized fashion becomes cumbersome, if not nearly impossible. The solution is to divide these operations over distributed data gathering and processing centers.

We are dealing now with a distributed system, and performing clustering and centralized operations over distributed data raises additional problems. Two of the most important problems are network transmission limitations and the security of transmitted data. A distributed system also requires distributed processing and clustering methods and algorithms.

In [63], da Silva et al. make an overview of the clustering algorithms used for distributed data mining, but with focus on applications in sensor networks. They also stress the importance of these clustering algorithms for distributed data retrieval and propose an algorithm for privacy-preserving clustering. Similar algorithms are required for AMI in the context of real-time distributed data processing, but these algorithms will be more permissive than the ones needed for sensor networks. Issues like limited power supply, limited power capability, asynchronous network topology which are present in the sensor network, have almost none, or very little impact in the Advanced Metering Infrastructure. Because of the sensitivity of the data transmitted between nodes, a greater emphasize should be placed on the confidentiality and integrity of data.

Working with data from distributed sources is an extensive new research area and it can motivate further studies and analysis, beyond the scope of this thesis.

5.5 First experiences using real data

In order to test the anonymization script, we have used data gathered from 5000 smart meters, during one month period. Also, we have built MATLAB scripts to verify the claims presented in Section 5.4.3. The MATLAB scripts can be found in Appendix B.

The first step is to anonymize the original smart meter's serial numbers. This is done as a batch job on a Linux machine by using the anon6gz.py python script. New features have been added to the anonymization script and the most important ones are: data consistency check, gunzip files support and a dictionary containing raw and anonymized values for improved speed.

The second step is to load the anonymized data files into MATLAB and run the prepared scripts on them. The script called dataparser.m reads the input files and creates a structure called 'sm' for every unique smart meter serial number contained by those files. In this way the overhead caused by the repetition of the serial number on the every line of the original file is eliminated. Additional space is saved by converting the date from a YYYYMMDDhhmmss format to a SDN (Serial Date Number) format which expresses every date as a float number representing the number of days which passed since 01-January-0000. For example the date 2011-10-15 20:33:45 in SDN format is 734791.856770833370000. MATLAB has internal functions which can easily make the transformation between SDN format and readable strings, in conventional time format. For every smart meter two arrays are created. The first one contains all the readings' timestamps in SDN format. The second array contains the readings' index at the specific timestamp. Additional fields can be added to the original data structure depending on what needs to be analysed (e.g. information about decreasing indexes, information about double recordings, arrays containing the power profile, etc.).

Next we present the results obtained by running the scripts which check for decreasing indexes and double recordings on the sample containing 5,000 smart meters. The results can be seen in Table 5.5.

As observed in Table 5.5, the percentage of smart meters which exhibit decreasing indexes is relatively high (40%) so we decided to follow this track and to further analyze them. We have found that the number of decreasing indexes per individual smart meter varies from 1 to a maximum of 22 (exhibited by only one smart meter - 0.02%). The decreasing value is usually small (under 1 kWh), but the large number of smart meters on which these events occur indicates an important avenue for future investigation. Figure 5.2 contains a graph representing the smart meters which exhibit decreasing indexes. The number of double recordings is relatively small and the recordings which

Smart meters	Number	Percent
Total	5000	100%
with decreasing indexes	2019	40%
with double recordings	321	6%
with decreasing indexes and double recordings	115	2%

Table 5.5: Smart meters with decreasing indexes and/or double recordings

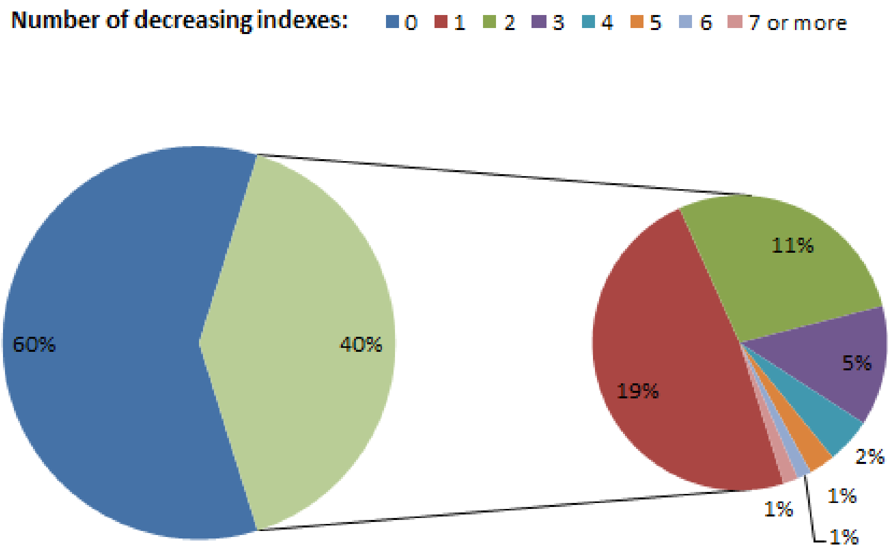

Figure 5.2: Smart meters exhibiting decreasing indexes

are repeated are actually duplicates (they have even the same index) and can be caused, for example, by repeated transmissions.

Daily mean power consumption profiles have been created for all the 5,000 smart meters contained in the sample. Figure 5.3 represents a graph of all these daily mean power profiles. It can be observed that a large number of power profiles are situated under the 25 kWh limit, but what is of interest for us at this point are the outliers. Some power profiles exhibit a strange downward slope at the beginning of the day or strange spikes which are well out of the normal profile. These oddities may be caused by corrupted data and explaining them and finding their cause is an issue for investigation that is not covered in this thesis.

As mentioned before, power consumption profiles can be used to infer information

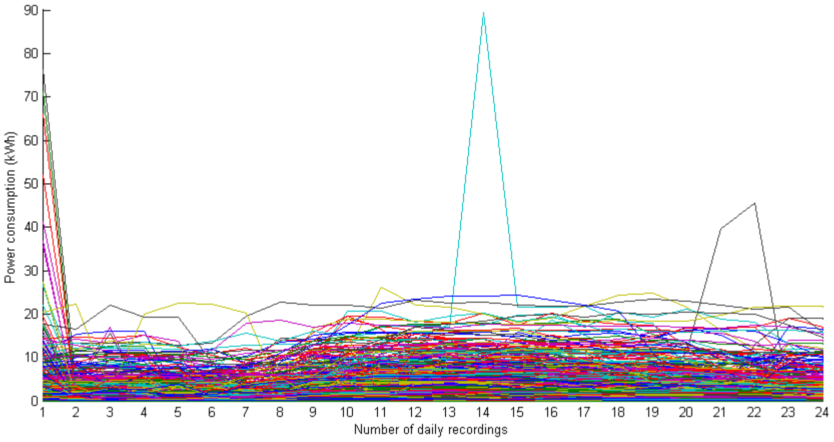

Figure 5.3: Daily mean power profiles for 5000 smart meters

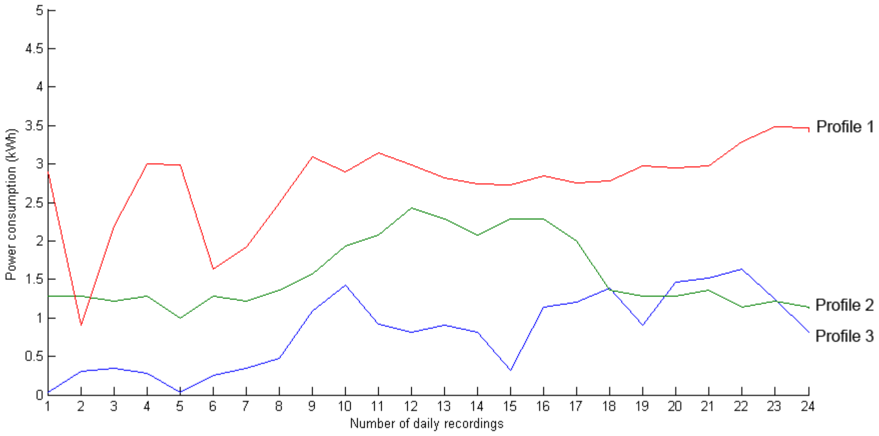

Figure 5.4: Daily mean power profiles for 3 smart meters

about what kind of activity is conducted in the building where the smart meter is installed. We have randomly chosen three mean daily power profiles which can be observed in Figure 5.4. Profile number 1 and profile number 2 may define two office locations. We observe a higher consumption during office hours (9:00 to 17:00) for profile number 2, while profile number 1 exhibits an almost constant high consumption for 2/3 of the day. Profile number 3 may define a domestic user which is a high consumer during morning, lunch and dinner, and cheaper energy cost in the evening may offer the possibility to use the washing machine or the dishwasher.

In conclusion we would like to say that these are some preliminary results based on our first experiences upon using real data, and these results can motivate further studies and analysis, beyond the scope of this thesis.

6

Conclusions and future work

In this research thesis we covered multiple aspects regarding the new Advanced Metering Infrastructure. Building the new Smart Grid is an ongoing process in many European countries. We thus emphasised that deploying the Advanced Metering Infrastructure requires the interaction and collaboration of specialists from different research and industry domains, mainly the EE domain and ICT domain.

We begun by presenting an overview of the important research regarding the Advanced Metering Infrastructure, from both the Electrical Engineering domain and the Information and Communication Technology domain, emphasizing cross-domain research.

We studied in more detail an important component of the Advanced Metering Infrastructure, the *smart meters*. We developed two scenarios where a skilled adversary may affect fundamental properties of the electrical grid by controlling a number of smart meters. By complementing and building on related research, we showed how problems in the distribution network may affect grid stability. The first scenario is presented from a theoretical perspective, and even though the necessary prerequisites of the scenario have been discussed in literature, the implications and limitations of the scenario are outlined here. The second scenario is studied in more detail, and a simulation is performed on a small *power island* to show feasibility. The conclusion of these two scenarios is that the ON/OFF feature should be designed and implemented with care and following a good security protocol. Also for a safe utilization of this feature, a large-scale turn OFF command should not be placed in the hands of only one entity, and should require the shared consent of several parties.

With the current push for massive installation of smart meters as well as a continuous development of their capabilities, it is only a question of time before the infrastructure is attacked. This asks for a developing and deploying process which is done with security in mind, because adding security features at a later stage would be difficult and costly. The security features development must involve researchers and engineers from both the Information and Communication Technology domain and the Electrical Engineering domain, because they must cover the individual threats and the shared ones.

Information provided by the smart meters is another subject we have covered. The use of real data in the research process is indispensable, since the outcome of the studies is rarely a theoretical model and it is more oriented towards real-life utilization.

Obtaining real data for studies is cumbersome, because data owners (electrical companies) are very concerned about the privacy of their customers. To simplify the process of data release, we proposed an anonymization solution for the information concerning customer identification, such as the serial numbers of smart meters or customer's ID numbers. This solution is based on a cryptographic procedure and provides consistency over different traces.

We continued by making a survey on how data from smart meters can be used to develop detection and mitigation techniques for attacks and energy fraud attempts, and we present some preliminary results obtained on first-time experiences involving real data. We covered only the off-line centralized data analysis case, and we believe that extending this and adding distributed data sources and real-time analysis is an important avenue for future research.

The main goal of our research is to look at the problems from an interdisciplinary point of view, considering both issues related to computer security and the electrical power domain. The smart grid straddles both these two domains and expertise on both areas is necessary to develop successful mitigation strategies.

Bibliography

[1] Observ'ER, Worldwide energy production from renewable energy sources, Stats and Figures Series (2010).
URL http://www.energies-renouvelables.org/observer/html/inventaire/pdf/12e-inventaire-Chap01-Eng.pdf

[2] A. Breidthardt, German government wants nuclear exit by 2022 at latest (May 2011).
URL http://uk.reuters.com/article/2011/05/30/us-germany-nuclear-idUKTRE74Q2P120110530

[3] SmartGrids – European Technology Platform (Jun. 2011).
URL http://www.smartgrids.eu/?q=node/163

[4] European Commission, European SmartGrids technology platform: Vision and strategy for Europe's electricity networks of the future (Apr. 2006).
URL http://ec.europa.eu/research/energy/pdf/smartgrids_en.pdf

[5] T. Goodspeed, Extracting Keys from Second Generation Zigbee Chips, in: Black Hat USA, Las Vegas, Nevada, 2009.

[6] S. McLaughlin, D. Podkuiko, P. McDaniel, Energy Theft in the Advanced Metering Infrastructure, in: Proceedings of the 4th Workshop on Critical Information Infrastructures Security (CRITIS), 2009.

[7] J. Fan, J. Xu, M. H. Ammar, S. B. Moon, Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme, Computer Networks 46 (2) (2004) 253 – 272.
URL http://www.sciencedirect.com/science/article/pii/S1389128604001197

[8] DLMS User Association, DLMS/COSEM Architecture and Protocols (2010).
URL http://www.dlms.com/documentation/index.html

[9] S. Keemink, B. Roos, Security Analysis of Dutch smart metering systems (2008).
URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.6314&rep=rep1&type=pdf`

[10] KEMA Consulting, P3 Companion Standard Dutch Smart Meter Requirements (2010).
URL `http://www.energiened.nl/_upload/bestellingen/publicaties/286_P3Dutch%20Smart%20Meter%20%20v2.1%20final%20P3.pdf`

[11] E. R. D. France, PLC G3 Profile Specification (2009).
URL `http://www.maxim-ic.com/products/powerline/pdfs/G3-PLC-Physical-Layer-Specification.pdf`

[12] T. Richardson, ZigBee the Wireless Standard for Tomorrow's Smart Grid (2010).
URL `http://intelligrid.epri.com/Smart_Grid_Information_Sharing_Calls/2010/100128/epri_ZigBee_presentation_richardson_012810.pdf`

[13] SCE&G, Underground vs. Overhead Power Lines (2011).
URL `http://www.sceg.com/NR/rdonlyres/465E6534-2FFB-4069-BF84-81465AEEF887/0/%20Undergroundvs.pdf,`

[14] A. Pabla, Electric Power Distribution, Tata McGraw-Hill, 2004.

[15] Y. Liu, P. Ning, M. Reiter, False data injection attacks against state estimation in electric power grids, in: Proceedings of the 16th ACM conference on Computer and communications security, Chicago, Illinois, 2009.

[16] H. Markiewicz, A. Klajn, Standard EN 50160 - voltage characteristics in public distribution systems (2004).

[17] B. Franken, V. Ajodhia, K. Petrov, K. Keller, C. Müller, Regulation of Voltage Quality, in: 9th International Conference "Electric Power, Quality and Utilisation", Barcelona, 2007.

[18] W. H. Kersting, Distribution System Modeling and Analysis, CRC Press, 2002.

[19] N. Mithulananthan, M. M. A. Salama, C. A. Canizares, J. Reeve, Distribution system voltage regulation and var compensation for different static load models, in: International Journal of Electrical Engineering, vo1.37, no. 4, pp.384-395, 2000.

[20] T. Ackermann, G. Andersson, L. Söder, Distributed generation: a definition, Electric Power Systems Research 57 (3) (2001) 195 – 204.
URL `http://www.sciencedirect.com/science/article/pii/S0378779601001018`

[21] G. W. Arnold, W. K. Reder, Why Building the Smart Grid Will be a Long-Term Project (2011).
URL `http://smartgrid.ieee.org/news-smart-grid-newsletter/4700-why-building-the-smart-grid-will-be-a-long-term-project`

[22] R. R. R. Barbosa, A. Pras, Intrusion detection in SCADA networks, in: Proceedings of the Mechanisms for autonomous management of networks and services, and 4th international conference on Autonomous infrastructure, management and security, AIMS'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 163–166.
URL `http://portal.acm.org/citation.cfm?id=1875873.1875903`

[23] H. Christiansson, E. Luiijf, Creating a European SCADA security testbed, in: E. Goetz, S. Shenoi (Eds.), Critical Infrastructure Protection, Vol. 253 of IFIP International Federation for Information Processing, Springer Boston, 2007, pp. 237–247.
URL `http://dx.doi.org/10.1007/978-0-387-75462-8_17`

[24] E. Johansson, T. Sommestad, M. Ekstedt, Issues of Cyber Security in SCADA Systems. On the Importance of Awareness., in: 20th International Conference on Electricity Distribution, 2009.

[25] H. Sandberg, A. Teixeira, K. H. Johansson, On Security Indices for State Estimators in Power Networks, in: Preprints of the First Workshop on Secure Control Systems, Stockholm, Sweden, 2010.

[26] A. Bose, Smart transmission grid applications and their supporting infrastructure, Smart Grid, IEEE Transactions on 1 (1) (2010) 11 –19.

[27] G. Ericsson, Toward a framework for managing information security for an electric power utility CIGRÉ experiences, Power Delivery, IEEE Transactions on 22 (3) (2007) 1461–1469.

[28] G. Ericsson, Management of information security for an electric power utility – on security domains and use of ISO/IEC17799 standard, IEEE Transactions on Power Delivery 20 (2) (2005) 683–690.

[29] G. Ericsson, Information security for electric power utilities (EPUs) – CIGRÉ developments on frameworks, risk assessment, and technology, Power Delivery, IEEE Transactions on 24 (3) (2009) 1174–1181.

[30] Y. Wang, D. Gu, J. Xu, H. Du, Hacking risk analysis of web trojan in electric power system, Web Information Systems and Mining, International Conference on (2009) 510–514.

[31] N. Falliere, L. O. Murchu, E. Chien, W32.Stuxnet Dossier (2011).
URL `http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf`

[32] P. H. Gleick, Water and terrorism (2006).
URL `http://www.pacinst.org/reports/water_terrorism.pdf`

[33] EU Forward, Forward: Managing Emerging Threats in ICT Infrastructures (2008).
URL `http://www.ict-forward.eu`

[34] J. R. D. G. Xuan Jin, John Bigham, C. Phillips, Anomaly detection in electricity cyber infrastructures, in: Proceedings of the International Workshop on Complex Networks and Infrastructure Protection (CNIP-06), 2006.

[35] D. G. Xuan Jin, John Bigham, C. Phillips, Test data for anomaly detection in electricity infrastructures, in: International Journal of Critical Infrastructures, Vol. 2, 2006, pp. 396 –411.

[36] D. Bergman, D. Jin, J. Juen, N. Tanaka, C. Gunter, A. Wright, Nonintrusive load-shed verification, Pervasive Computing, IEEE 10 (1) (2011) 49–57.

[37] K. Samarakoon, J. Ekanayake, Demand side primary frequency response support through smart meter control, in: Proceedings of the 44th International Universities Power Engineering Conference (UPEC), 2009, pp. 1–5.

[38] DynamicDemand (June 2011).
URL `http://www.dynamicdemand.co.uk/grid.htm`

[39] R. Berthier, W. Sanders, H. Khurana, Intrusion detection for advanced metering infrastructures: Requirements and architectural directions, in: First IEEE International Conference on Smart Grid Communications (SmartGridComm), 2010, pp. 350–355.

[40] A. Metke, R. Ekl, Security technology for smart grid networks, Smart Grid, IEEE Transactions on 1 (1) (2010) 99–107.

[41] J. Zerbst, M. Schaefer, I. Rinta-Jouppi, Zone principles as cyber security architecture element for smart grids, in: IEEE PES Conference on Innovative Smart Grid Technologies Europe (ISGT Europe), 2010, pp. 1–8.

[42] M. Carpenter, T. Goodspeed, B. Singletary, E. Skoudis, J. Wright, Advanced Metering Infrastructure Attack Methodology, `http://inguardians.com/pubs/AMI_Attack_Methodology.pdf` (2009).

[43] F. Boroomand, A. Fereidunian, M. Zamani, M. Amozegar, H. Jamalabadi, H. Nasrollahi, M. Moghimi, H. Lesani, C. Lucas, Cyber security for smart grid: A human-automation interaction framework, in: Innovative Smart Grid Technologies Conference Europe (ISGT Europe), 2010 IEEE PES, 2010, pp. 1 –6.

[44] J. Zerbst, M. Schaefer, I. Rinta-Jouppi, Zone principles as cyber security architecture element for smart grids, in: Innovative Smart Grid Technologies Conference Europe (ISGT Europe), 2010 IEEE PES, 2010, pp. 1 –8.

[45] A. Metke, R. Ekl, Security technology for smart grid networks, Smart Grid, IEEE Transactions on 1 (1) (2010) 99 –107.

[46] H.-H. So, S. Kwok, E. Lam, K.-S. Lui, Zero-configuration identity-based signcryption scheme for smart grid, in: First IEEE International Conference on Smart Grid Communications (SmartGridComm), 2010, pp. 321–326.

[47] F. Li, B. Luo, P. Liu, Secure information aggregation for smart grids using homomorphic encryption, in: First IEEE International Conference on Smart Grid Communications (SmartGridComm), 2010, pp. 327–332.

[48] E. Hayden, There is No SMART in Smart grid without secure and reliable communication (2010).
URL `http://www.verizonbusiness.com/resources/whitepapers/wp_no-smart-in-smart-grid-without-secure-comms_en_xg.pdf`

[49] S. Mclaughlin, D. Podkuiko, A. Delozier, S. Miadzvezhanka, P. Mcdaniel, Embedded firmware diversity for smart electric meters, in: Proceedings of the 5th USENIX Workshop on Hot Topics in Security (HotSec 2010), Washington DC., 2010.

[50] C. Efthymiou, G. Kalogridis, Smart grid privacy via anonymization of smart metering data, in: Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, 2010, pp. 238 –243.

[51] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, D. Irwin, Private memoirs of a smart meter, in: 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys 2010), Zurich, Switzerland, 2010.
URL `http://www.cs.umass.edu/~kevinfu/papers/molina-markham-buildsys10.pdf`

[52] G. Kalogridis, C. Efthymiou, S. Denic, T. Lewis, R. Cepeda, Privacy for smart meters: Towards undetectable appliance load signatures, in: Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, 2010, pp. 232 –237.

[53] R. Anderson, S. Fuloria, Who controls the off switch?, in: Proceedings of the IEEE SmartGridComm, 2010.

[54] J. F. Prada, M. D. Ilic, The value of reliability in power systems - pricing operating reserves - (1999).
URL `http://web.mit.edu/energylab/www/pubs/el99-005wp.pdf`

[55] PowerWorld Corporation (Jun. 2011).
URL `http://www.powerworld.com/products/simulator.asp`

[56] S. Meliopoulos, G. Cokkinides, R. Huang, E. Farantatos, S. Choi, Y. Lee, X. Yu, Smart grid infrastructure for distribution systems and applications, in: System Sciences (HICSS), 2011 44th Hawaii International Conference on, 2011, pp. 1 –11.

[57] Z. Fadlullah, M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, Y. Nozaki, Toward intelligent machine-to-machine communications in smart grid, Communications Magazine, IEEE 49 (4) (2011) 60 –65.

[58] Federal Information Processing Standards, Announcing the ADVANCED ENCRYPTION STANDARD (AES) (2001).
URL `http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf`

[59] E. W. S. dos Angelos, O. R. Saavedra, O. A. C. Cortés, A. N. de Souza, Detection and identification of abnormalities in customer consumptions in power distribution systems, Power Delivery, IEEE Transactions on 26 (4) (2011) 2436 –2442.

[60] F. Fabris, L. Margoto, F. Varejao, Novel approaches for detecting frauds in energy consumption, in: Network and System Security, 2009. NSS '09. Third International Conference on, 2009, pp. 546 –551.

[61] J. Cabral, J. Pinto, A. Pinto, Fraud detection system for high and low voltage electricity consumers based on data mining, in: Power Energy Society General Meeting, 2009. PES '09. IEEE, 2009, pp. 1 –5.

[62] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, A. Wu, An efficient k-means clustering algorithm: analysis and implementation, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (7) (2002) 881 –892.

[63] J. C. da Silva, C. Giannella, R. Bhargava, H. Kargupta, M. Klusch, Distributed data mining and agents, Engineering Applications of Artificial Intelligence 18 (7) (2005) 791 – 807.
URL `http://www.sciencedirect.com/science/article/pii/S095219760500076X`

A

Appendix A

A.1 PowerWorld Simulator Line Properties

Buses	Length (km)	Impedance (ohm/km)
1 – 2	0.5	0.05
2 – 3	0.5	0.05
3 – 4	0.5	0.05
5 – 6	0.5	0.05
6 – 7	0.5	0.05

Table A.1: Underground Line Properties

Buses	Length (km)	Impedance (ohm/km)
1 – 2	0.5	0.15
2 – 3	0.5	0.15
3 – 4	0.5	0.15
5 – 6	0.5	0.15
6 – 7	0.5	0.15

Table A.2: Overhead Line properties

Load	2	3	4	5	6	7
Min(kWh)	2	7	2	25	25	8
Max(kWh)	67	63	65	50	49	61
Avg(kWh)	16	24	17	40	40	22

Table A.3: Loads on every bus

B

Appendix B

B.1 Anonymization procedure code for the smart meter serial number or customer ID - JAVA

```
import java.io.*;
import javax.crypto.*;
import javax.crypto.spec.*;
import java.math.BigInteger;
import java.util.*;

public class AESwithPass {

 public static byte[] encrypt (byte[] ceva) throws Exception {
  /**
    * The AES encrypton algorithm using ECB.
    * Here the encryption key can be modified to any 128, 192 or 256 length key.
    */
  String seqkey = ”abcdefghijklmnop”;
  byte[] raw = seqkey.getBytes(”UTF-8”);
  SecretKeySpec key = new SecretKeySpec(raw,”AES”);
  Cipher cipher = Cipher.getInstance(”AES/ECB/NOPADDING”,”SunJCE”);
  //Cipher cipher = Cipher.getInstance(”AES”);// longer time to execute
  cipher.init(Cipher.ENCRYPT_MODE, key);
  return cipher.doFinal(ceva);
 }

 public static String anonymize(String id, String pad) throws Exception {
  /**
    * The anonimization technnique used in CryptoPAn.
    */
  String todo;
 StringBuilder forxor = new StringBuilder();
 forxor.append(’0’);
 //start padding and encrypting
 for(int i = 1; i < id.length(); i++){
  todo = id.substring(0,i)+pad.substring(0, 127−i);
  BigInteger conv = new BigInteger(todo,2);
  byte[] tod = Arrays.copyOf(conv.toByteArray(),16);
  byte[] encrypted = encrypt(tod);
  BigInteger plm = new BigInteger(encrypted);
  String xul = plm.toString(2);
  forxor.append((char)xul.charAt(xul.length()−1));
  }
 return forxor.toString();
}

public static String zeroPadding(String bin, int len){
 /**
   * This function gets as input a value as string,
   * returning the same value padded with zeros to the specified length
   */
 StringBuilder zero = new StringBuilder();
 for (int toprepend=len−bin.length(); toprepend>0; toprepend−−) {
  zero.append(’0’);
  }
 zero.append(bin);
 return zero.toString();
}

public static void main(String[] args) throws Exception {
 /**
   *The Padding KEY can be modified here (it is encrypted with AES as in CryptoPAn);
   *Data is read/written from/to file (matching the 3 semicolumn separated values
   *meter id, date and consumption index);
   *The meter id is converted to binary, padded with zeros, encrypted and
   *xored with the original value, resulting the anonymized ID;
   */

 long start = System.currentTimeMillis();

 String seqpad = ”1234567890ABCDEF”;
 byte[] pad = seqpad.getBytes(”UTF−8”);
 byte[] padding = encrypt(pad);
 BigInteger tz = new BigInteger(padding);
 String bitPadding = zeroPadding(tz.toString(2), 128);

 FileReader rstream = new FileReader(”consumptionC.txt”);
 BufferedReader plain = new BufferedReader(rstream);
 FileWriter fstream = new FileWriter(”EncryptedC.txt”);
 BufferedWriter crypt = new BufferedWriter(fstream);
```

```
  String line;
  while ((line = plain.readLine()) != null) {
   String datavalue[] = line.split(”;”);
   if (datavalue.length!=3) {
    break;
    }
   BigInteger n = new BigInteger(datavalue[0]);
   String date = datavalue[1];
   String consumption = datavalue [2];

   String meterid = zeroPadding(n.toString(2), 60);
   String ency = anonymize(meterid, bitPadding);
   Long xored =Long.parseLong(meterid,2) ^ Long.parseLong(ency.toString(),2);

   String testL = xored.toString();
   if (testL.length()>18) {
    testL = testL.substring(testL.length()-18);
    }
   else if (testL.length()<18) {
    testL = zeroPadding(testL, 18);
    }
   crypt.write(testL);
   crypt.write(”;”);
   crypt.write(date);
   crypt.write(”;”);
   crypt.write(consumption);
   crypt.newLine();
  }
  crypt.close();
  plain.close();

  long end = System.currentTimeMillis();
  float time = new Float(end-start);//TimeUnit.MILLISECONDS.toSeconds(end-start);
  System.out.println(”Execution time was ”+time/1000+” seconds.”);
  System.out.println(”DONE.”);
 }
}
```

B.2 Anonymization procedure code for the smart meter serial number or customer ID - Python

```
#! /usr/bin/python
#This module requires also the installation of two additional modules :
#python-bitstring -> http://code.google.com/p/python-bitstring/
#pycrypto -> http://www.pycrypto.org
#
#prefix preserver , quick method
#v6. added dictionary for improved anonymization speed
#please see dictionary size variables at lines 18-20
import os
import sys
import re
import random
import time
import gzip
from bitstring import BitArray,BitStream
from Crypto.Cipher import AES

slen=18 #serial length - this is fixed to 18, don't modify
plen=10 #prefix length - this can vary between 1 and 18
        #for no prefix please use plen = 18

oldser={} #dictionary {raw_serial:anonymized_serial,...}
dictsize=500 #how many elements should be kept in the dictionary
dictquota=dictsize/10 #the quota that is freed when the dictionary size
                            #exceeds dictsize

lc=0    #line counter
lerr=0  #number of lines with errors
#encryption key is read from an external file
ckey=''  #first half as encryption key (128 bits - 16 chars)
pad=''  #second half as padding (128 bits - 16 chars)

def cryptokey(kline):
    key=kline;
    if len(key)==32:
        ckey=key[0:16]  #first half as encryption key (128 bits - 16 chars)
        pad=key[16:32]  #second half as padding (128 bits - 16 chars)
        return ckey,pad
    else:
        print "The encryption key must have 32 alphanumeric characters \
                and be on the first line of the file containing it. \
                Please check your key file."
        sys.exit(2)
```

```
def ser2bin(ser): #converts serial number from int to binary
    try:
        res=BitArray(uint=int(ser), length=60)
        return res
    except ValueError:
        res=0

def str2bin(str): #converts string to binary
    return BitArray(bytes=str,length=128)

def sec128(plain): #encrypts the pad before using
    e = AES.new(ckey, AES.MODE_ECB)
    secret=e.encrypt(plain)
    return secret

def anonymizepref(serial,bepad): #anonymization procedure
    global slen,plen
    otp=BitArray(uint=0, length=1) #the one time pad;f0=0 constant function
    bser=ser2bin(serial)    #converting meter int serial to binary
    if bser is None:
        return ''
    for i in range(1,len(bser)):
        spadded=bser[0:i-1] #create a0a1..ai-1
        spadded.append(bepad[0:128-i+1])    #pad with bits from encrypted pad


        spad=spadded.tobytes() #converting from bits to bytes for encryption
                                                        #R(P(a0a1...ai-1),k)

        eres=str2bin(sec128(spad))  #encryption R(P(a0a1...ai-1),k)

        otp.append(bser[i-1:i]^eres[(len(eres)-1):len(eres)])    #XORing immediatly
    sserial=otp
    short_ser=str(sserial.int)  #convert anonymized serial to string form

    if len(short_ser)==plen:
        return short_ser

    if len(short_ser)<plen:
        return short_ser.zfill(plen-len(short_ser))
               #padding left with necessary zero

    if len(short_ser)>plen:
        return short_ser[(-plen):]
               #returning only the last plen decimal digits

def anonymizesuf(serial,bepad): #anonymization procedure
    global slen,plen
    otp=BitArray(uint=0, length=1) #the one time pad;f0=0 constant function
    bser=ser2bin(serial)    #converting meter int serial to binary
    if bser is None:
        return ''
    for i in range(1,len(bser)):
        spadded=bser[0:i-1] #create a0a1..ai-1
        spadded.append(bepad[0:128-i+1])
                #pad with bits from encrypted pad until 128bits

        spad=spadded.tobytes() #converting from bits to bytes for encryption
                                                                        #R(P(a0a1...ai-1),k)

        eres=str2bin(sec128(spad))  #encryption R(P(a0a1...ai-1),k)

        otp.append(bser[i-1:i]^eres[(len(eres)-1):len(eres)])
                #XORing immediatly
    sserial=otp
    short_ser=str(sserial.int)  #convert anonymized serial to string form

    if len(short_ser)==slen-plen:
        return short_ser

    if len(short_ser)<slen-plen:
        return short_ser.zfill((slen-plen)-len(short_ser))
                #padding left with necessary zero

    if len(short_ser)>(slen-plen):
        return short_ser[(plen-slen):]
                #returning only the last plen-slen decimal digits

def sliceandanonymize(i,bepad):
#does individual anonymization for prefix and suffix
    global slen,plen
    res=''
    pref=i[0:plen]
    suf=i[plen:slen]
    pa=anonymizepref(pref,bepad)
    if len(pa)==0:
        return res
    sa=anonymizesuf(suf,bepad)
    res=pa+sa
    return res

def anon2file(tfile,line,bepad): #line parsing
    global oldser
    global slen,plen
    m=re.split(';',line) #regular expression
    ss='' #ss will hold the new anonymized serial
    if len(m[0])!=slen:
        return 1
    if (m[0] in oldser): #uses the old anonymized serial if the serial repeats
        ss=oldser[m[0]]+';'+m[1]+';'+m[2]
        tfile.write(ss)
    else:
        ss=sliceandanonymize(m[0],bepad)
                #anonymizes the serial based on prefix and suffix length
        if (len(ss)!=slen):
            return 1 #checks the serial meters length

        oldser[m[0]]=ss #adds the new serial to dictionary
        ss=ss+';'+m[1]+';'+m[2]
        tfile.write(ss)
    return 0

def purgedict():
    global oldser
    global dictsize
    kk=len(oldser) #returns the number of keys in the dictionary
    #print kk
    if (kk>dictsize):
        for k in range(int(kk-dictsize+dictquota)):
            (k,v)=oldser.popitem()
    #kk=len(oldser) #returns the number of keys in the dictionary
    #print kk


def main():
    global lc #line counter
    global lerr #counts the lines with errors
    global ckey,pad
    if len(sys.argv)!=4:
        print "File usage: "+sys.argv[0]+" <input_file> <output_file> <key_file>"
        sys.exit(2)

    ts=time.time()  #used to count the anonymization time

    ofile=gzip.open(sys.argv[1],'rb') #the original file
    tfile=gzip.open(sys.argv[2],'wb') #the target file
    kfile=open(sys.argv[3],'r+') #the file containing the key
    (ckey,pad)=cryptokey(kfile.readline()) #checks the encryption key
    kfile.close()

    print "Encryption key is: "+ckey
    print "128bit pad is: "+pad

    bepad=str2bin(sec128(pad)) #encrypt pad before using

    #for line in ofile: #parsing the original file
    line=ofile.readline()

    while len(line)!=0:
        if(anon2file(tfile,line,bepad)==1):
            lerr+=1

        lc+=1
        if lc%1000==0:
            print "# of parsed lines: %d" % lc
            purgedict()
        line=ofile.readline()

    ofile.close()
    tfile.close()
    te=time.time()  #used to count the anonymization time

    print "All done! Output written to file: "+sys.argv[2]
    print "%d lines have been read." %lc
    print "%d line(s) didn't follow the required format." %lerr
    print "Anonymization took %.6f seconds." % (te-ts)

main()
```

B.3 MATLAB scripts

```
%This script reads data from a smart meter file and returns a structure such as:
%   sm
%    |
%    |---.sn //smart meter serial number e.g. 000111222333444555
%    |
%    |---.ts //array of timestamps in datenum format
                         (number of days from 1-Jan-0000) e.g. [734454.354143519;...]
%    |                   see help datenum,datevec,datestr for more info
%    |---.index //array of smart meter index value   e.g. [123.56;122,75;...]
%    |

inputfn = input('Enter name of file:   ', 's'); %input file

if (exist('sm')==1) %adds to the current structure if it exists
    si=length(sm)+1;

elseif (exist('sm')==0) %creates a new structure if not
    si=1; %index of smart meters
    sm.sn=''; %smart meter serial number
    sm.ts=[];   %time stamp array
    sm.index=[];    %index array
    sm.di=[];    %array for decreasing indexes
    sm.dd=[];    %array for double recordings
    smmap=containers.Map();
```

```
%creates a new map for smart meter serial number -> position in structure
end

i=0;
fid=fopen(inputfn,'r'); %opens the file in readonly mode
nlines=0;% counting number of processed lines
tline = fgetl(fid);%skips the first line
tline = fgetl(fid);%string which needs regex
nlines=nlines+2;


while ischar(tline) %while is data on current line
   i=i+1; %increment smart meter structure index
   rline=regexp(tline,';','split'); %perform regular expression on line
   tsn=char(rline(1)); %interpret serial number
   tts=datenum(char(rline(2)),'yyyymmddHHMMSS'); %interpret date as datenum
   tin=str2double(rline(3)); %interpret index as float number

   if (si==1) %if this is the first recording
       sm(si).sn=tsn;   %add serial number
       sm(si).ts=[sm(si).ts;tts]; %add timestamp to current serial number
       sm(si).index=[sm(si).index;tin]; %add index to current serial number
       smmap(tsn)=si; %add smart meter sn to map
       si=si+1; %increment index in structure
   elseif (isKey(smmap,tsn))
%if the current serial number is equal to one already in the data structure
       pos=smmap(tsn); %retrieve the position in the structure from the map
       sm(pos).ts=[sm(pos).ts;tts]; %add the current timestamp
       sm(pos).index=[sm(pos).index;tin]; %add the current index
   else
%create a new recording in the data structure if smart meter is not present
       sm(si).sn=tsn; %add serial number
       sm(si).ts=[sm(si).ts;tts]; %add timestamp
       sm(si).index=[sm(si).index;tin]; %add index
       smmap(tsn)=si; %add smart meter sn to map
       si=si+1;
   end

   tline = fgetl(fid); %read next line
   nlines=nlines+1; %count the line
   if (mod(nlines,1000)==0)
    sprintf('%d lines processed',nlines)
%print message when lines is multiple of 1000
   end
end

%uses the sm structure created by dataparser.m to sort the recordings
%ascending based on timestamp
```

```
for i=1:length(sm)
    TS=[sm(i).ts';sm(i).index'];
    SS = sortrows(TS.',1).';
    sm(i).ts=SS(1,:).';
    sm(i).index=SS(2,:).';

end

clear i TS SS

clear fid i inputfn rline si tin tline tsn tts nlines pos

%uses the sm structure created by dataparser.m to check for decreasing index value
%(possible energy fraud) for the same SM
smc=0;
ti=0;

for i=1:5000
    %length(sm)
    for j=1:length(sm(i).ts)-1
        %tf=isequal(sm(i).ts(j,1:6),sm(i).ts(j+1,1:6));
      if (sm(i).ts(j)<sm(i).ts(j+1))
        if (sm(i).index(j)>sm(i).index(j+1)) %if we have decreasing index
            sm(i).di=[sm(i).di;j]; %record the event
            if (i>ti)
                smc=smc+1;
                ti=i;
            end
        end
      end
    end
end
 sprintf('%d smart meters had decreasing indexes\n
To see these smart meters and the indexes please run seedecindexes.m',smc)



clear i j tf smc ti

% uses the sm structure to print all the smart meters which have decreasing
% indexes

for i=1:length(sm)
    if (~isempty(sm(i).di)) %check if we have recordings for current sm
        sprintf('Smart meter #%s has %d decreasing indexes.', \
        sm(i).sn,length(sm(i).di))
        inputq = input('Do you want to see them? (y/n/q)', 's');
        if strcmp(inputq,'y')
        for j=1:length(sm(i).di)
        sprintf('%s -> %.5f \n',datestr(sm(i).ts(sm(i).di(j)),\
        'yyyy-mm-dd HH:MM:SS'),sm(i).index(sm(i).di(j)))
        sprintf('%s -> %.5f \n',datestr(sm(i).ts(sm(i).di(j)+1), \
        'yyyy-mm-dd HH:MM:SS'),sm(i).index(sm(i).di(j)+1))
        end
        elseif strcmp(inputq,'q')
            clear i j inputq
            break
        end
        end
    end


clear i j inputq

%uses the sm structure created by dataparser.m to check for double
%recordings (double dates) for the same SM
smc=0;
ti=0;

for i=1:5000
    %length(sm)
    for j=1:length(sm(i).ts)-1
        tf=isequal(sm(i).ts(j),sm(i).ts(j+1));
        %check for equal succesive timestamps
        if (tf==1) %if found
            sm(i).dd=[sm(i).dd;j]; %record the event
            if (i>ti)
                smc=smc+1;
                ti=i;
            end

        end
    end
end
sprintf('%d smart meters had double recordings\n
To see these smart meters and the recordings please run seedoublerec.m',smc)

clear i j tf smc ti

% uses the sm structure to print all the smart meters which have double
% recordings

for i=1:length(sm)
    if (~isempty(sm(i).dd)) %check if we have recordings for current sm
        sprintf('Smart meter #%s has %d double recordings.', \
        sm(i).sn,length(sm(i).dd))
        inputq = input('Do you want to see them? (y/n/q)', 's');
        if strcmp(inputq,'y')
```

```
            for j=1:length(sm(i).dd)
            sprintf('%s -> %.5f \n',datestr(sm(i).ts(sm(i).dd(j)), \
'yyyy-mm-dd HH:MM:SS'),sm(i).index(sm(i).dd(j)))
            sprintf('%s -> %.5f \n',datestr(sm(i).ts(sm(i).dd(j)+1), \
'yyyy-mm-dd HH:MM:SS'),sm(i).index(sm(i).dd(j)+1))
            end
        elseif strcmp(inputq,'q')
            clear i j inputq
            break
        end
        end
    end


clear i j inputq

%uses the sm structure created by dataparser.m to create a power
%consumption profile for a single Smart Meter

%structure to keep the power profiles
clear smp
smp.sn=''; %smart meter serial number
smp.ts=[]; %time stamps
smp.index=[]; % consumption index
smp.pwprofd=[]; %daily power consumption profile, ld recordings/day
smp.pwprofdm=[];%mean daily profile
%smp.pwprofm=[]; %monthly power consumption profile, lm recordings/month
%smp.pwprofmm=[];%mean monthly profile
%end structure to keep the power profiles

%create data for general power profile
for i=1:length(sm)
    smp(i).sn=sm(i).sn;
    smp(i).ts=sm(i).ts;
    smp(i).index(1)=0;%the first recording should be always 0
    for j=1:length(sm(i).ts)-1
      smp(i).index=[smp(i).index;sm(i).index(j+1)-sm(i).index(j)];

    end
end


clear i j tf p tt ld lm ly ti pf pi

%uses smp structure to create daily consumption profiles for all smart
%meters

eps=0; %we set maximum 24 recordings per day
%epsilon - how many additional recordings to add each day
```

```
maxrecperday=0; %check for maximum recordings per day
ld=24+eps;  %daily
lm=31*ld;   %monthly
%create daily power profiles
for i=1:5000
    if mod(i,100)==0
        sprintf('%d smart meters processed.',i)
    end
    %length(smp)
    ti(24)=0;
     pi=1; %initial pivot
    for j=1:length(smp(i).ts)-1
        pf=j; %end pivot
        if (floor(smp(i).ts(j))==floor(smp(i).ts(j+1)))
            %if (ti(str2double(datestr(smp(i).ts(j),'HH'))+1)==0)
                ti(str2double(datestr(smp(i).ts(j),'HH'))+1)=smp(i).index(j);
            %end
            %if (floor(smp(i).ts(j+1))==23)
            %     if (ti(str2double(datestr(smp(i).ts(j),'HH'))+2)==0)
                ti(str2double(datestr(smp(i).ts(j),'HH'))+2)=smp(i).index(j+1);
            %     end
            %end
            pf=j+1;

        else
            if maxrecperday<pf-pi %store the maximum number of recordings per day
                maxrecperday=pf-pi+1 %for future use
                i
            end
            %ti=smp(i).index(pi:pf); %copy the indexes on last day
            pi=pf+1;
            %ti=[ti.',NaN(1,ld-length(ti))];
            smp(i).pwprofd=[smp(i).pwprofd;ti];

            ti(24)=0;

        end
    end
end

%create daily mean consumption profile
for i=1:5000
    %length(smp)
    smp(i).pwprofdm=mean(smp(i).pwprofd);
end

clear i j tf p tt ld lm ly ti pf pi
```