THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Enhancing Privacy and Security in the Advanced Metering Infrastructure

VALENTIN TUDOR

Division of Networks and Systems
Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2015

Enhancing Privacy and Security in the Advanced Metering Infrastructure
Valentin Tudor

Technical report 130L
ISSN 1652-876X
Department of Computer Science and Engineering
Division of Networks and Systems
Chalmers University of Technology
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31-772 10 00

Author e-mail: tudor@chalmers.se

Printed by Chalmers Reproservice
Göteborg, Sweden 2015

Enhancing Privacy and Security in the Advanced Metering Infrastructure

Valentin Tudor

Division of Networks and Systems, Chalmers University of Technology

ABSTRACT

Cyberphysical systems introduce computing and communication devices within different sectors of society generating large quantities of data. Information extracted from these data gives better understanding of these sectors and the systems within them. There are many ways in which the information extracted from these data can be harnessed. For example, monitoring these devices is important in order to keep account of their correct functionality, especially in critical infrastructures, and the aforementioned data can be used to construct and enhance such monitoring systems.

Processing these data is not always easy, first due to their complexity and also due to the sensitive personal information that might be inferred from them. The processing must be done with care in order to preserve the privacy of the users whose behavior generates the data in question. Different methods and technologies have been proposed in order to preserve the privacy of these personal data and they are tailored to the same data that they are protecting. Because of this, it is important to study the influence of the characteristics of these data on the effectiveness of the privacy preserving solutions employed.

The work in this thesis focuses on the Advanced Metering Infrastructure (AMI) in the smart electrical grid and it has two goals. The first one is to study the characteristics of the AMI data and the thesis investigates the effect of these characteristics on a number of privacy enhancing technologies which are proposed for AMI data. The second one is to work towards an Intrusion Detection System (IDS) for the AMI, and the thesis presents an important IDS module which processes encrypted traffic and is able to infer the different commands run between AMI devices, without decrypting the traffic or accessing the sensitive data.

Keywords: Advanced Metering Infrastructure, data privacy, communication security, intrusion detection

Acknowledgments

First, I would like to express my gratitude to my supervisors, Dr. Marina Papatriantafilou and Dr. Magnus Almgren for their constant guidance, encouragement and understanding.

My work has been partially supported by the European Comission Seventh Framework Programme (FP7/2007-2013) through the FP7-SEC-285477-CRISALIS project and the collaboration framework of Chalmers Energy Area of Advance.

I would like to thank the current and former members of the Networks and Systems division for the wonderful and friendly working environment. Thank you Ali, Anders, Aras, Aljoscha, Andreas, Bapi, Beshr, Daniel, Elad, Elena, Erland, Eugenio, Farnaz, Georgios, Ioannis, Iosif, Ivan, Katerina, Mustafa, Nasser, Nhan, Olaf, Oscar, Paul, Peter L., Philippas, Pierre, Thomas, Tomas, Viktor, Vincenzo and Zhang. Thank you for all the discussions, support and feedback you have provided so far. I would also like to thank Anneli, Eva, Jonna, Marianne, Peter H., Rolf, Rebecca and Tiina for their great administrative support.

Last, but not least, I would like to thank my family for their constant encouragement, support and understanding over all these years — *Mulţumesc mult!* My greatest thanks goes to my lovely fianceé, Simona, for her endless love and support. Thank you, I could not have gone so far without you!

Valentin Tudor
Göteborg, April 2015

List of Appended Papers

I **Valentin Tudor**, Magnus Almgren, Marina Papatriantafilou,
"Analysis of the impact of data granularity on privacy for the smart grid",
in *Proceedings of the 12th ACM workshop on privacy in the electronic society (WPES 2013)*, Berlin, Germany, November 4, 2013, pp. 61—70.

II **Valentin Tudor**, Magnus Almgren, Marina Papatriantafilou,
"A study on data de-pseudonymization in the smart grid",
in *Proceedings of the eighth European Workshop on Systems Security (EuroSec 2015)*, Bordeaux, France, April 21, 2015, Article No. 2.

III **Valentin Tudor**, Magnus Almgren, Marina Papatriantafilou,
"Harnessing the unknown in Advanced Metering Infrastructure traffic",
in *Proceedings of the 30th ACM/SIGAPP Symposium On Applied Computing (SAC 2015)*, Salamanca, Spain, April 13—17, 2015, pp. 2204—2211.

Contents

List of Figures

Part I

INTRODUCTION

Chapter 1

Introduction

The trend in today's society is to collect more and more data from automatic processes and human behavior [25]. These data give us better understanding of the different areas from where they are collected and also allow us to improve the individual systems which generate these data, making them more efficient.

Different sectors in our society that were previously communication-wise independent are becoming increasingly interconnected. This is possible with the help of distributed sensor networks, and some of these networks have gained direct Internet connection making data produced by them available to even larger networks. The simplest example is the mobile communication sector, where devices evolved from providing only telephone services to offering multiple sensor-based applications (GPS, accelerometers) and also direct connection to the Internet. The transportation area is experiencing similar upgrades, as some car manufacturers already provide Internet communication capabilities to some of their models. It is envisioned that Internet connected cars will become standard in the future [21]. The energy sector is another example where new Information and Communications Technology (ICT) devices bring advantages to the process of monitoring the production and delivery of energy. As part of the concept of intelligent homes [9], which is based on these devices, it is envisioned to make the dwellers more informed about how they consume

electrical energy and what their environmental footprint is. At the same time, this information can also be leveraged to better adjust the balance between energy production and consumption and help in the integration of local renewable energy sources [43].

The equipment from the above-mentioned sectors and many more adjacent ones produce large quantities of data that need to be processed and transformed into valuable information. This information can then be used to improve different aspects of these sectors, which in turn have a big impact in improving the quality of peoples' lives. One of the main benefits of processing these large quantities of data is to keep account of the correct functionality of the distributed equipment that serve these sectors. Monitoring the behavior of the equipment is extremely important, as some of them are deployed in infrastructures which are critical for the function of our society, and their failure or malicious utilization may cause serious damage to property and even to human lives [18, 33, 40]. This raises the need for new monitoring methods and solutions in order to better control the capabilities and functionalities of these new devices.

Although their benefits are significant, the large quantities of data collected also raise serious challenges due to the sensitive information that can be inferred from them. In many of these areas, data produced directly or indirectly by human beings raise privacy concerns in both the policy and the academic environments. Historically, one of the most privacy sensitive sectors was the medical one: patients' data, especially in electronic format, need to be stored and processed following strict procedures [38]. The vehicular sector has similar concerns, as data produced by the cars can be used to track the whereabouts of the drivers and their driving style, which can be privacy invasive [50]. Similarly, data produced by devices in different sections of the electrical network can be used to infer information about the customers' lifestyles [34]. All these examples point towards the importance of studying the privacy issues raised by the large quantities of data produced in all the areas aforementioned. Due to the similarities in the data producing process and of the human presence, solutions found for *monitoring equipment* and *privacy related* issues in one of these areas are many times applicable to others.

This thesis studies these two critical aspects in a particular section of the electrical energy sector: data privacy and equipment behavior monitoring in the *Advanced Metering Infrastructure (AMI)*. The AMI is one of the components of the electrical network where the electrical engineering (EE) meets the information and communications technology (ICT). The work in the thesis explores the large data produced in the AMI and it is two-folded. The first goal is to study the characteristics of the data that is collected and this thesis investigates the effect of these characteristics on a number of privacy enhancing technologies which are proposed for AMI data. The second goal is to work towards an Intrusion Detection System for the AMI environment, and this thesis presents an important IDS module which processes encrypted traffic and is able to infer the different commands run between AMI devices, without decrypting the traffic or accessing the sensitive data.

The rest of this introductory chapter is structured as follows: in Section 1.1 we present the general aspects of the Advanced Metering Infrastructure, we outline the privacy issues raised by the data collection and also the need for an Intrusion Detection System monitoring the communication network. In Section 1.2 we formulate our research questions related to the aforementioned aspects and we present our research contributions. Section 1.3 gives a summary of the appended publications and we conclude this introductory chapter with Section 1.4.

1.1 Security and privacy aspects of the Advanced Metering Infrastructure (AMI)

1.1.1 The smart electrical grid and the AMI

During the last decade, the electrical grid has undergone a number of changes in its transition to the so-called *smart grid* [52]. There is no clear definition, but the concept of smart grids can be summarized as done by the European Commission Directorate General for Research 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" [52].

Some of the main changes are related to the installation of new ICT-based equipment in different parts of the classical electrical grid. The electrical grid has three main sections: the generation section (electricity is produced in power plants), the transmission section (electricity is transported over high-voltage lines) and the distribution section (in the end it is delivered to the consumers). Each section needs to be carefully monitored and managed in order to ensure the proper functionality of the equipment and overall stability of the electrical network.

The generation and transmission sections are monitored and managed by the *Supervisory Control and Data Aquisition system (SCADA)*. Although the concept of SCADA predates the smart grid, this type of control system also face some changes with installation of new devices and upgrade of the historical legacy systems with new ones that use commercial ICT equipment and operating systems [27]. Important changes have also taken place in the distribution section where the classical electrical energy meters are replaced with new ones, called *smart meters* [10]. Compared to their predecessors, the smart meters have two-way communication with the central system and their extended functionalities can produce additional data that can be an important source of information for the system operator. These devices, together with the communication network that connects them, form the *Advanced Metering Infrastructure* (AMI) [5]. The AMI is relatively new and it brings together elements from electrical engineering and information technology. The transition of the distribution network is not complete and the AMI deployments might differ from country to country and even from area to area depending on the local rules, regulations and implementations [51].

There are a number of improvements expected to be achieved with the transition to the smart grid, two of the most important being a better management of the consumption of electrical energy and the integration of renewable energy sources [47]. As electrical energy is one of the most important resources for today's society, the smart grid is, or will be, closely connected with other areas,

as it manages the delivery of energy to other important sectors of our society. For example, a close interaction with the vehicular domain is possible, especially since the electrical vehicles will have an active role in the electrical grid as energy sources during peak periods [12, 48]. The AMI is a critical component of the electrical energy sector which is important for our nowadays society; therefore, studying its properties and particularities, finding solutions for its inherent problems and developing tools for the AMI is extremely important from a research perspective.

1.1.2 Large datasets - multiple possibilities

The size of the smart grid is considerable, with devices installed in key points in the electrical network. It is estimated that data produced by smart grid equipment will be considerable and the size of the smart grid will be larger than the size of the Internet[1]. It is expected that these data will play a key role in the development of the smart grid, and analyzing and building applications on these data will contribute towards improving electrical grid stability and energy consumption efficiency.

In this thesis we focus mainly on data produced by equipment in the AMI. As described above, the main actor here is the new type of electrical energy meter, called the smart meter. The deployment of such meters is not complete, but it is envisioned that each household will be served by at least one smart meter with communication capabilities [17] - thus the network serving the AMI environment will have a considerable size.

The smart meters are capable of providing fine-grained data that reflect the electrical energy consumed at the household level, but also data about the quality of the electrical power delivered [5]. Billing of electrical energy consumed by the household is done using the kWh energy consumption value. The billing is usually done over long periods of time (i.e. 1-3 months), but the fine-grained energy consumption data can also be used for differential tariffs or even for customer re-imbursements when renewable energy is produced locally [47].

[1] `http://news.cnet.com/8301-11128_3-10241102-54.html`

The smart meters are capable of providing data which show the instantaneous values of voltage, current, active and reactive power. These values can be used for grid operation purposes and they can be very useful for a so-called *low-voltage SCADA* system [45], especially when local renewable energy sources are integrated into the electrical grid. Compared with data required for billing, grid operational data need to be collected very often (i.e. every few minutes) in order to give an accurate overview of the electrical distribution network. Efthymiou and Kalogridis [14] use the term *high-frequency (HF) data* for data which are used for grid operational purposes and *low-frequency (LF) data* for data used for billing purposes. We will keep the same definitions for these two types of data throughout this work.

Smart metering data allow for a number of additional applications such as peak energy consumption shaving [20], short-term energy consumption forecast [49], prevention of energy-related fraud [7, 35], securing critical infrastructures [1, 23, 24] and also educating the consumers towards efficient energy usage [36]. These applications show the importance and also the complexity of the information that can be extracted from these smart metering data, making this type of data a valuable source of information and an important component in the process of improving the electrical grid.

1.1.3 Privacy aspects of large datasets

Whenever large quantities of data are produced it is important that sensitive information is properly protected. When talking about data privacy, the medical sector comes to mind [42] as patients' information needs to be protected, especially if it is stored in an electronic format. However, fine-grained data produced in the AMI can also be sensitive: previous research [34, 41] showed that it can be used to infer information about the lifestyle of the people living at the premises, such as electrical equipment usage patterns and even presence or absence from the premises. At the time of writing, there is no specific European Directive that covers smart metering data in particular so this type of data falls under the incidence of the EU Data Protection Directive 95/46/EC [13]. As a

consequence, the deployment of AMIs can be significantly slowed down when privacy preservation is not guaranteed by law [11]. Currently, the EU Smart Grid Task Force[2] offers documents containing information and recommendation for smart grid investors and data controllers.

Privacy related issues raised by the sensitive data collected in the Advanced Metering Infrastructure caught the attention of the research community which proposed different methods to preserve and/or enhance the privacy of the smart metering data. These solutions can be data specific and rely on data anonymization [14] or data aggreation [32], or they can even rely on technical solutions that require energy storage equipment [29, 30]. Data aggregation solutions based on homomorphic cryptography [32, 34, 54] are proposed as privacy-preserving solutions. These solutions are successful in preserving the privacy of the households involved, but depending on how the aggregation is performed, some important information might be lost, thus narrowing the applications where these data can be used.

We return to the two types of AMI data mentioned earlier: high-frequency (HF) data used for grid operational purposes and low-frequency (LF) data used for billing. Efthymiou and Kalogridis [14] consider billing data to be privacy neutral, as they are collected seldom (LF data), thus showing aggregated information about the energy consumption process over the time period considered. Also, for a correct billing, these data need to be attributable to a specific household. On the other hand, grid operational data need to be collected often (HF data) and they might show detailed information about the households' lifestyle. This type of data needs to be collected in an anonymized fashion, thus breaking the connection with the real identity of its producer. This can be acomplished with the help of third party entities which are responsible of collecting the sensitive data, anonymizing it and then delivering it to the data beneficiary [3, 14, 53]. Separating these two types of data is also proposed by Borges et al. [4] in a solution based on anonymity networks where a customer uses different identities for transmitting the billing and the grid-operational data.

[2] `http://ec.europa.eu/energy/en/topics/markets-and-consumers/smart-grids-and-meters`

Storing billing and operational data under different identities seems like a viable privacy-preserving solution, but recent studies have shown that only a fraction of the households are protected, while for others it is possible to break the anonymity and tie together the pseudo-identities of the data producers [6, 28]. Studying the characteristics of the AMI data and especially how they influence the efficiency of the privacy-preserving methods [15] will help towards improving these methods, efficient collection of AMI data and also privacy-preserving smart grid services [16, 22]. We further describe this problem and our research contribution in Research Question 1 (RQ1) in Section 1.2.1.

1.1.4 Towards building an Intrusion Detection System (IDS)

Monitoring ICT systems and their communication network is important in order to detect equipment malfunction or even malicious activities, especially when the monitored equipment is part of a critical infrastructure. Specialized systems, called *Intrusion Detection Systems (IDS:s)* are responsible for monitoring and analyzing system events in order to provide warnings regarding suspicious activity. IDS:s [46] have three logical components: the sensors (data collectors), the analyzers (data processors) and the user interface (presenting the information to the operator). Depending on its location and on the type of data analyzed, the IDS:s can be host-based (monitoring the events in a single host), network-based (monitoring the network traffic) or distributed/hybrid (mix between host-based and network based). Historically, the IDS:s were mainly used on computer systems and networks. With the penetration of ICT systems in new sectors such as vehicular communication and smart grid, new IDS:s need to be developed in order to fulfill the new monitoring requirements.

As previously mentioned, the smart grid is responsible for managing the production, transmission and the delivery of electrical energy, so by its nature it becomes a critical infrastructure. It is important to monitor the devices themselves and also their communication network in order to detect user or equipment misbehavior and even attacks. The possible effects of attacks against crit-

ical infrastructures became clearer with the discovery of Stuxnet [18] and its capacity to infect and destroy industrial equipment. Fortunately there has been no major attacks against the electrical grid recorded so far, but the research community is working for identifying vulnerabilities that require mitigation. McLaughlin et al. [37] and Grochocki et al. [23] present possible attack scenarios for the AMI, starting from potential attacker goals covering denial of service, energy fraud, and even targeted disconnect of electrical services. For example, a distributed denial of service (DDoS) attack might be conducted by instructing a large number of smart meters to flood a specific power station [40]. This might cause problems in the ICT equipment and it might even affect the energy supply in the area serviced by the targeted power station. Also, AMI might be the target for economic fraud. There are a few recent recorded cases of AMI energy fraud in the wild in which some customers have tampered with the smart meters in order to lower their electricity bill. KrebsonSecurity [31] reported a case of meter fraud involving many customers from Puerto Rico, which may have cost the utility several hundreds of millions of dollars. A similar case happened in Malta [33] where the authorities discovered that at least $1,000$ smart meters had been tampered with to record lower electricity consumption levels. Around 10% of the total local generation of electricity was affected, causing a loss of approximately $41 million in 2012 alone.

These events that were reported in the wild and also the concerns raised by the research community express the need for monitoring solutions for the smart grid in general and for the AMI in particular. Carpenter et al. [8] present a number of vulnerabilities that exist in devices in the AMI together with a possible attack methodology against devices in the AMI. Subsequently, Foreman and Gurugubelli [19] present the attack surface of the AMI with respect to hardware and network configurations, protocols, and software. Both of these studies are a good starting point for developing IDS services for this environment.

Raciti and Nadjm-Tehrani [44] express the need and present a model for a host-based IDS which focuses on the detection of anomalies inside the AMI devices. They build a module based on this model and they test four possible types of attacks against the internals of the smart meters: data manipulation

attack, recalibration attack (changing registers' values), reset attack (deleting the records regarding consumed energy) and sleep mode attack (the meter is put into sleep mode and the energy consumed is not registered). All these attacks, if not detected can lead to serious economic losses, as previously presented.

The idea of a distributed monitoring solution for the smart grid is advocated by Zhang et al. [55] with multiple IDS nodes deployed at different points in the AMI enviroments and also by Grochocki et al. [23] who proposed a distributed IDS model that can scale with the size of the AMI network as well as being able to monitor traffic between peers. Due to the lack of known attack signatures, Mitchell and Chen [39] stress that a behavior (anomaly) based IDS is preferred in this environment over a signature-based one. Detecting events and understanding the correct behavior of AMI equipment requires processing of some of the large data produced here. An important source of information comes from monitoring the traffic between devices [26] but this process might become cumbersome, especially when the AMI traffic is encrypted in order to maintain confidentiality [2].

All these unaddressed issues open important research questions, as the need for monitoring the system needs to be balanced with the one for confidentiality. In traditional ICT systems, monitoring would reveal the user's interaction with the system such as detailed information about his web surfing habits. Monitoring the AMI, if not done with care, might reveal additional information about the lifestyle of the inhabited households, and it might even become more privacy invasive, as some aspects of the energy consumption data might still be unknown and out of the control of the customer. This shows that there is a need to develop monitoring solutions for the AMI that are also privacy preserving and the starting point for these monitoring tools relies in the study of the large data produced in the AMI. We further describe this problem and our research contribution in Research Question 2 (RQ2) in Section 1.2.2.

1.2 Research questions and contributions

Our research is based on the analysis of the large data produced in the AMI and has two aims. The first is related to the characteristics of the data collected in the AMI and it is focused on understanding their effect on the different anonymization methods that were previously proposed. The second focuses on the process of monitoring the communication between AMI devices in the case when the traffic is encrypted in order to preserve the privacy of the customers' data and also the confidentiality of the communication. We briefly formulate our two research questions here and we present them each in detail in the following subsections:

- RQ1 - What is the effect of the characteristics of AMI datasets on the efficiency of privacy enhancing methods?
- RQ2 - Is it possible to balance the need for confidentiality and customers' data privacy with the need to monitor the communication network?

In the following we present each of these research questions, our contributions as well as a definition of our research aims.

1.2.1 RQ1: AMI data characteristics and anonymization efficiency

As previously mentioned, privacy in general and also in the AMI in particular is important as the lack of it has a direct effect on the individuals. Data gathered in the AMI in particular can be used to infer information about the living patterns of the residents. We investigate how characteristics of collected AMI datasets can influence the efficacy of previously proposed anonymization techniques for the AMI.

In particular, we study the case of two types of AMI data, one used mainly for billing and which should be safe with regard to privacy concerns (called Low-Frequency (LF) data [14]) and one that could be used for grid operations but which is privacy sensitive (called High-Frequency (HF) data [14]). The

former is stored using the real identity of the smart meter that generated them while the latter is stored under a pseudonym to enhance the privacy of the data.

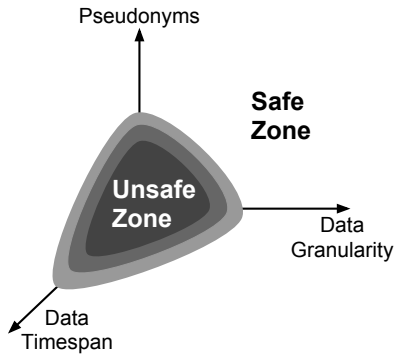

Figure 1.1: Characteristics of AMI datasets

Figure 1.1 (proposed by us in Paper I) presents the three characteristics that we focus on: the usage of *pseudonyms* in the process of reporting/storing data for the same smart meter, the *timespan* of data stored by the utility provider under the same pseudonym for the same smart meter and the *granularity* of reported/stored data.

Reporting high-frequency data under different pseudonyms and making sure that connections between pseudonyms are extremely hard to find and/or known only by a trusted third party [4, 14, 32] have been proposed earlier in the research literature. Using one pseudonym can be useful, if the connection between this pseudonym and the real smart meter identity is secret, but reporting or storing data from the same smart meter under different pseudonyms for shorter timespans can further enhance the privacy of the AMI data [4].

The timespan of data stored under the same pseudonym/identity is also very important, because longer periods of stored data for a smart meter (under the same pseudonym) help to create an accurate power consumption profile which can be used to extract information about the energy consumption pattern. This information can be used in connecting together the consumption profiles of the same smart meter that are stored under different pseudonyms.

The last characteristic taken into consideration is the granularity of reported and stored data. Low-frequency data must be reported using the exact consumption values for accurate billing and to prevent fraud. The question is whether the reported high-frequency data can be altered in a minor way such that the modification will not affect the grid operation, but making it hard to identify each customer uniquely by, for example, making the data from different customers more uniform.

We investigate these three characteristics in two de-anonymization scenarios, based on two separate adversary models. In the first scenario we study how the granularity and the timespan of the data stored are influencing the adversary's capability of linking together two datasets produced by the same smart meters. One dataset contains HF smart metering data stored with the help of pseudonyms, while the other dataset contains LF smart metering data stored under the real identity of the smart meter. In the second scenario we investigate an adversary whose goal is to use consumption features extracted from the datasets containing HF data in order to link together the different pseudonyms used for the same smart meter. Starting from a previously proposed de-pseudonymization method we develop a simpler one and we investigate for both methods the effect of the dataset size and the seasonality of the data on the adversary's success rate.

Our methodology and results can be used by electricity companies to better understand the properties of their smart metering datasets and the conditions under which such datasets can be released to third parties.

1.2.2 RQ2: Monitoring AMI equipment by analyzing encrypted traffic

Due to security and privacy concerns, the communication between the AMI devices is encrypted, making it more secure against malicious third parties but also obscuring the ability of the network owner to detect any misbehaving user or equipment. We are investigating how to balance the need for confidentiality with the need to monitor the AMI which is important in the context of de-

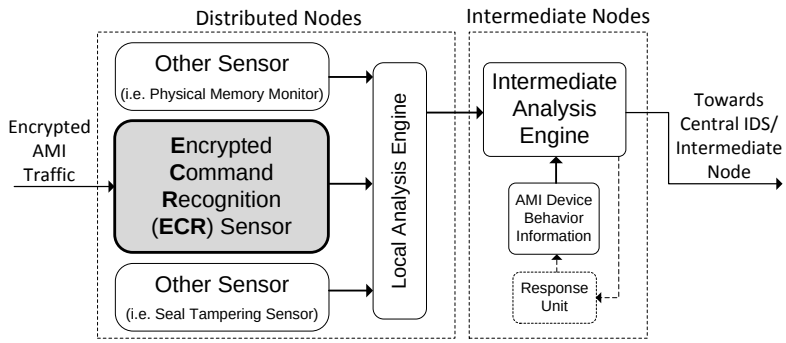

Figure 1.2: The *ECR* sensor in an AMI Intrusion Detection System

veloping intrusion detection solutions for the AMI. We develop one important component for an AMI Intrusion Detection System (IDS), which we call the *Encrypted Command Recognition (ECR)* sensor (depicted in Figure 1.2 and proposed by us in Paper III). ECR's purpose is to accurately determine the individual commands that are passed in an encrypted or hard to parse form in an AMI communication network. This is done in a privacy-preserving fashion, without decrypting the traffic and without accessing the sensitive data transmitted in the AMI communication network.

We analyze the characteristics of the devices in the AMI and based on their functionality, communication patterns and protocols we extract a set of features that characterize the communication protocols used. Compared with the classical ICT domain, where the human being is the communication initiator and modeler through his interaction with the computer systems, in the AMI the communication is mainly initiated automatically, by the devices themselves. Even though a human operator sometimes would manually query AMI devices, the largest volume of the communication follows certain patterns [26], as the automatic readings will take the majority of the traffic exchanged while the maintenance commands might be concentrated in a fixed time period. These extracted features are used to create a classifier with the help of supervised learning and

we study its efficiency on traffic captured in a realistic AMI testbed. We focus on two AMI protocols, one which is encrypted and one which is difficult to parse due to its proprietary implementation.

The ECR module can become an important component of a distributed IDS for the AMI environment and it will give the operator a better view on the status of his network and help in the early detection of possible misbehaviors and even attacks.

1.3 Summary of appended publications

1.3.1 Paper I: Analysis of the impact of data granularity on privacy for the smart grid

In the first paper we investigate RQ1 in detail. We focus on the energy consumption traces that smart meters generate and especially on the risk of being able to identify individual customers given a large dataset of these traces. We present a first step towards an analytical framework that can be used to estimate this risk based on the uniqueness of the customers. We make a formalization of the problem of de-anonymization by matching low-frequency and high-frequency smart metering datasets and we build an adversary model for this problem.

We compare the theoretical model estimation results with evaluation results obtained from a large dataset of smart metering energy consumption data captured in a live enviroment. Our findings show that two characteristics of AMI data (data timespan and granularity) play a big role in the AMI data anonymization process. We show how these characteristics can be tuned in order to reduce the number of linked identities, thus enhancing customers' privacy.

1.3.2 Paper II: A study on data de-pseudonymization in the smart grid

In the second paper we continue the investigation of RQ1 and we study the influence of AMI data characteristics on the efficiency of privacy-preserving methods. Starting from an already proposed de-pseudonymization scenario in the literature we develop a simpler method based on statistical information extracted from the consumption patterns that helps to better model a potential adversary that wants to tie together datasets belonging to the same household.

We show that the number of re-identified households is dependent on the size of the dataset and the period where the pseudonyms are constant and not changed. In the context of the smart grid, results will even vary based on the season when the dataset is captured. Knowing that relative simple changes in the data collection procedure may significantly increase the resistance to de-anonymization attacks will help future AMI deployments.

1.3.3 Paper III: Harnessing the unknown in Advanced Metering Infrastructure traffic

In the third paper we investigate RQ2 and we study how to balance the need for confidentiality with the need to monitor the traffic between AMI equipment. More specifically, we develop one important component for an AMI IDS, which can accurately determine the individual commands (but not their content) exchanged between AMI devices even when they are sent over an encrypted channel or in a protocol that the IDS cannot parse.

We propose a methodology and a number of features which summarize traffic characteristics. We demonstrate, for two AMI protocols, that a set of commands can be identified with high accuracy by using only the previously mentioned information. This component is important as it can be used by an IDS for the AMI in order to maintain a behavior model of each equipment instance in an AMI network. This will give the operator a better view on the status of his network and help in the early detection of possible attacks and misbehaviors.

1.4 Conclusions and future work

It is clear that the penetration of ICT devices in various sectors and their communication capabilities will produce large quantities of data which, after being processed and transformed into valuable information, can be used to improve the respective sectors. Similarly, it is almost unquestionable that the smart grid will produce more and more data reflecting the electrical energy consumed and also the status of the devices in the electrical grid. Harnessing and processing these large quantities of data will make the electrical grid more resilient to faults, provide a better balance between the production and the consumption of energy, but by their nature, these datasets also raise privacy concerns. We propose the first steps towards a framework to analyze the effect of the smart metering data characteristics on the efficiency of some of the curently proposed privacy-preserving methods. We investigate three main characteristics of smart metering data: the granularity of the data reported, its timespan and the number of pseudonyms used for reporting these data. Our results show that they can play a big role in a three-way balance towards obtaining better customer anonymity. We will continue this research path by expanding the adversary model towards one that obtains only partial information, based on the similarities of energy consumption patterns.

We also investigated how to balance the need for confidentiality with the need to monitor the traffic between AMI equipment. We proposed a methodology for an encrypted command recognition component which can be used in an AMI IDS. We showed how statistical information can be used to correctly identify AMI commands, even when they are sent over an encrypted channel or embedded in a protocol that is hard to parse. This is especially useful in AMI environments where the operator wants to keep the traffic encrypted, both for security reasons (protection against the use of sensitive commands, such as the remote on-off switch) and for customers' privacy.

Creating a universal traffic analysis tool for all possible AMI deployments is very challenging, due to the different types of protocols and network topologies that exist. We expect that the features identified for the two AMI protocols

covered in this study and the analysis of their performance in identifying the different AMI commands will give useful insight on the internal characteristics of these types of protocols. We expect that our study will facilitate the analysis of other proprietary and/or encrypted AMI protocols as a step to build an IDS to protect this critical infrastructure.

We presented the first steps towards a behavior-based Intrusion Detection System for the AMI environment that uses this module and additional information in order to maintain a behavior model of each equipment instance in an AMI network. This will give the operator a better view on the status of his network and help in the early detection of possible attacks and misbehaviors.

Although our investigation of security and privacy issues is focused on the AMI environment, our findings are not limited only to this specific area. These findings can be useful also in similar large scale areas such as sensor networks and vehicular networks which share many common characteristics through the large data produced by the similar ICT components.

Bibliography

[1] R. Berthier and W. H. Sanders. Specification-based intrusion detection for advanced metering infrastructures. In *2011 IEEE 17th Pacific Rim International Symposium on Dependable Computing (PRDC)*, pages 184–193, December 2011.

[2] R. Berthier, D. I. Urbina, A. A. Cárdenas, M. Guerrero, U. Herberg, J. G. Jetcheva, D. Mashima, J. H. Huh, and R. B. Bobba. On the practicality of detecting anomalies with encrypted traffic in AMI. In *Proceedings of the IEEE Conference on Smart Grid Communications (SmartGridComm)*, 2014.

[3] J.-M. Bohli, C. Sorge, and O. Ugus. A privacy model for smart metering. In *Communications Workshops (ICC), 2010 IEEE International Conference on*, pages 1–5, May 2010.

[4] F. Borges, L. A. Martucci, and M. Mühlhaüser. Analysis of privacy-enhancing protocols based on anonymity networks. In *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, pages 378–383. IEEE, 2012.

[5] R.E. Brown. Impact of smart grid on distribution system design. In *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, pages 1–4, July 2008.

[6] E. Buchmann, K. Böhm, T. Burghardt, and S. Kessler. Re-identification of smart meter data. *Personal and ubiquitous computing*, 17(4):653–662, 2013.

[7] J. E. Cabral, J. O. P. Pinto, and A. M. A. C. 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.

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

[9] M. Chan, D. Estève, C. Escriba, and E. Campo. A review of smart homes - present state and future challenges. *Computer methods and programs in biomedicine*, 91(1):55–81, 2008.

[10] Federal Energy Regulatory Commission. Assessment of demand response and advanced metering. 2008.

[11] C. Cuijpers and B.-J. Koops. Smart metering and privacy in Europe: Lessons from the Dutch case. In *European data protection: coming of age*. Springer, 2013.

[12] S. Deilami, A.S. Masoum, P.S. Moses, and M.A.S. Masoum. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. *Smart Grid, IEEE Transactions on*, 2(3):456–467, Sept. 2011.

[13] EU Directive. 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the EC*, 23:6, 1995.

[14] C. Efthymiou and G. Kalogridis. Smart grid privacy via anonymization of smart metering data. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 238 –243, Oct. 2010.

[15] G. Eibl and D. Engel. Influence of data granularity on smart meter privacy. *Smart Grid, IEEE Transactions on*, 6(2):930–939, March 2015.

[16] Z. Erkin and T. Veugen. Privacy enhanced personal services for smart grids. In *Proceedings of the 2nd Workshop on Smart Energy Grid Security*, SEGS '14, pages 7–12, New York, NY, USA, 2014. ACM.

[17] EU Commision. Report: Benchmarking smart metering deployment in the EU-27 with a focus on electricity /* COM/2014/0356 final */. `http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52014DC0356&from=EN`, July 2014. [last visited March 2015].

[18] N. Falliere, L. O. Murchu, and E. Chien. W32.Stuxnet Dossier. *White paper, Symantec Corp., Security Response*, 5, 2011.

[19] J. C. Foreman and D. Gurugubelli. Identifying the cyber attack surface of the Advanced Metering Infrastructure. *The Electricity Journal*, 28(1):94–103, 2015.

[20] G. Georgiadis and M. Papatriantafilou. Dealing with storage without forecasts in smart grids: Problem transformation and online scheduling algorithm. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14. ACM, 2014.

[21] M. Gerla and L. Kleinrock. Vehicular networks and the future of the mobile internet. *Computer Networks*, 55(2):457–469, 2011.

[22] Y. Gong, Y. Cai, Y. Guo, and Y. Fang. A privacy-preserving scheme for incentive-based demand response in the smart grid. *IEEE Transactions on Smart Grid*, PP, 2015.

[23] D. Grochocki, J. H. Huh, R. Berthier, R. Bobba, W. H. Sanders, A. A. Cárdenas, and J. G. Jetcheva. AMI threats, intrusion detection requirements and deployment recommendations. In *Smart Grid Communications (SmartGridComm), IEEE Third International Conference on*, pages 395–400, 2012.

[24] V. Gulisano, M. Almgren, and M. Papatriantafilou. METIS: a two-tier intrusion detection system for advanced metering infrastructures. In *Proceedings of the 5th international conference on Future energy systems*, pages 211–212. ACM, 2014.

[25] S. Gunelius. The Data Explosion in 2014 Minute by Minute - Infographic. `http://goo.gl/drrrxG`, July 2014. [last visited March 2015].

[26] M. Hoeve. Detecting intrusions in encrypted control traffic. In *Proceedings of the First ACM Workshop on Smart Energy Grid Security*, SEGS '13, pages 23–28. ACM, 2013.

[27] V. M. Igure, S. A. Laughter, and R. D. Williams. Security issues in SCADA networks. *Computers & Security*, 25(7):498–506, 2006.

[28] M. Jawurek, M. Johns, and K. Rieck. Smart metering de-pseudonymization. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 227–236. ACM, 2011.

[29] G. Kalogridis, R. Cepeda, S.Z. Denic, T. Lewis, and C. Efthymiou. Elecprivacy: Evaluating the privacy protection of electricity management algorithms. *Smart Grid, IEEE Transactions on*, 2(4):750–758, Dec. 2011.

[30] G. Kalogridis, C. Efthymiou, S.Z. Denic, T.A. Lewis, and R. Cepeda. Privacy for smart meters: Towards undetectable appliance load signatures. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 232–237, Oct. 2010.

[31] KrebsonSecurity. FBI: Smart Meter Hacks Likely to Spread. `http://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread/`, April 2012. [last visited March 2015].

[32] K. Kursawe, G. Danezis, and M. Kohlweiss. Privacy-friendly aggregation for the smart-grid. In *Privacy Enhancing Technologies*, pages 175–191. Springer, 2011.

[33] Malta Independent. Sparks fly over smart meter theft scandal. http://www.independent.com.mt/articles/2014-02-16/news/sparks-fly-over-smart-meter-theft-scandal-3968892934/, February 2014. [last visited March 2015].

[34] F.G. Mármol, C. Sorge, O. Ugus, and G.M. Pérez. Do not snoop my habits: preserving privacy in the smart grid. *Communications Magazine, IEEE*, 50(5): 166–172, May 2012.

[35] D. Mashima and A. A. Cárdenas. Evaluating electricity theft detectors in smart grid networks. In *Research in Attacks, Intrusions, and Defenses*, pages 210–229. Springer, 2012.

[36] K. Matsui, H. Ochiai, and Y. Yamagata. Feedback on electricity usage for home energy management: A social experiment in a local village of cold region. *Applied Energy*, 120(0), 2014.

[37] S. McLaughlin, D. Podkuiko, S. Miadzvezhanka, A. Delozier, and P. McDaniel. Multi-vendor penetration testing in the Advanced Metering Infrastructure. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 107–116. ACM, 2010.

[38] M. Meingast, T. Roosta, and S. Sastry. Security and privacy issues with health care information technology. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 5453–5458. IEEE, 2006.

[39] R. Mitchell and I.-R. Chen. Behavior-rule based intrusion detection systems for safety critical smart grid applications. *IEEE Transactions on Smart Grid*, 4(3):1254–1263, September 2013.

[40] Y. Mo, T. H.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli. Cyber–physical security of a smart grid infrastructure. *Proceedings of the IEEE*, 100(1):195–209, 2012.

[41] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin. Private memoirs of a smart meter. In *Proceedings of the second ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, BuildSys '10, pages 61–66, New York, NY, USA, 2010. ACM.

[42] T. Neubauer and J. Heurix. A methodology for the pseudonymization of medical data. *International Journal of Medical Informatics*, 80(3):190–204, 2011.

[43] C. W. Potter, A. Archambault, and K. Westrick. Building a smarter smart grid through better renewable energy information. In *Power Systems Conference and Exposition, 2009. PSCE'09. IEEE/PES*, pages 1–5. IEEE, 2009.

[44] M. Raciti and S. Nadjm-Tehrani. Embedded cyber-physical anomaly detection in smart meters. In *Critical Information Infrastructures Security*, volume 7722 of *Lecture Notes in Computer Science*, pages 34–45. 2013.

[45] S. Repo, D. Della Giustina, G. Ravera, L. Cremaschini, S. Zanini, J.M. Selga, and P. Jarventausta. Use case analysis of real-time low voltage network management. In *Innovative Smart Grid Technologies (ISGT Europe), 2011 2nd IEEE PES International Conference and Exhibition on*, pages 1–8, Dec. 2011.

[46] W. Stallings and L. Brown. *Computer Security Principles and Practice, Third edition*. Pearson Education, 2015.

[47] G. Strbac. Demand side management: Benefits and challenges. *Energy policy*, 36(12):4419–4426, 2008.

[48] W. Su, H. Eichi, W. Zeng, and M-Y. Chow. A survey on the electrification of transportation in a smart grid environment. *Industrial Informatics, IEEE Transactions on*, 8(1):1–10, Feb. 2012.

[49] J. W. Taylor. An evaluation of methods for very short-term load forecasting using minute-by-minute british data. *International Journal of Forecasting*, 24(4), 2008.

[50] T. Toledo, O. Musicant, and T. Lotan. In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transportation Research Part C: Emerging Technologies*, 16(3):320 – 331, 2008. Emerging Commercial Technologies.

[51] J. Torriti, M. G. Hassan, and M. Leach. Demand response experience in Europe: Policies, programmes and implementation. *Energy*, 35(4):1575–1583, 2010.

[52] SmartGrids – European Technology Platform `http://www.smartgrids.eu/?q=node/163`, June 2011. [last visited March 2015].

[53] S. Wang, L. Cui, J. Que, D-H. Choi, X. Jiang, S. Cheng, and L. Xie. A randomized response model for privacy preserving smart metering. *Smart Grid, IEEE Transactions on*, 3(3):1317 –1324, Sept. 2012.

[54] Y. Yan, Y. Qian, and H. Sharif. A secure data aggregation and dispatch scheme for home area networks in smart grid. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–6, Dec. 2011.

[55] Y. Zhang, L. Wang, W. Sun, R.C. Green, and M. Alam. Distributed intrusion detection system in a multi-layer network architecture of smart grids. *Smart Grid, IEEE Transactions on*, 2(4):796–808, 2011.