Karlsruhe Institute of Technology

Institut für Programmstrukturen und Datenorganisation

Lehrstuhl Prof. Dr.-Ing. Klemens Böhm

Privacy-Enhancing Methods for Time Series and their Impact on Electronic Markets

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

an der Fakultät für Informatik

des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Dipl.-Inform. Stephan Kessler

aus Karlsruhe

Datum der mündlichen Prüfung:	27. Mai 2015
Erster Gutachter:	Prof. Dr.-Ing. Klemens Böhm
Zweiter Gutachter:	Prof. Dr. rer. pol. Christof Weinhardt

Acknowledgments

Several people contributed significantly to this thesis. In this section, I would like to express my deep gratitude to these people.

First and foremost, I would like to thank my supervisor *Klemens Böhm* for his advice and continuous support. Without his help, this thesis would have been impossible. He already encouraged me to start doing research when I was still studying computer science and supported my application at the IME graduate school. His support continued during my time as a PhD student with constant and challenging reviews as well as fruitful discussions.

Erik Buchmann greatly contributed to the supervision: *Erik* did not only lay the foundation of my research direction, he also always helped me out when I got stuck and supported me intensively. It was always fun and inspiring to work with you *Erik* - thank you for your infinite patience and support.

I would also like to thank *Christoph Flath* for his help and great influence on the interdisciplinary part of my research. It was a great pleasure to work with you.

Besides the direct support from my supervisors, my colleagues at my chair and in my graduate school were also a continuous source of encouragement. Especially, I would like to thank *Patricia Iglesias*, *Fabian Keller* and *Philipp Ströhle*: You did not only contribute scientifically with lots of discussions, but also encouraged me to go on in difficult times. Thank you for being more than a colleague - for being my friend.

Likewise, I would like to thank the 'Expert Circle' (*Matthias Bracht*, *Clemens Heidinger* and *Heinz Herrmann*) for all the fun and good times we had visiting matches and discussing pros and cons of bets and soccer teams. In particular, I would like to thank you *Heinz* for having always an open door and an open ear, and for giving me many wise advices.

Last but not least I would like to thank my friends and family for always supporting me. *Christian* and *Alex* for being my friends for more than twenty years now. Thank you for encouraging me in the hard times and for celebrating with me the good times. *Magdalena* and *Daniel*, thank you for putting me back on solid ground again and again. Thank you *Uli* for your help and your 'plan' on the very last meters. To my parents, *Ingrid* and *Hans-Jürgen*, thank you for your strong support at anytime and for enabling me to fulfill my dreams.

Zusammenfassung des Inhalts

Der Themenkomplex Privatheit und Datenschutz rückt zunehmend in das Bewusstsein der Gesellschaft. Im Zuge dessen wurden nicht nur strengere Gesetze erlassen, sondern Privatheit spielt auch in der öffentlichen Wahrnehmung eine immer größere Rolle. Aufgrund einer Vielzahl neuer Technologien, werden auch immer mehr Zeitreihen persönlicher Daten gespeichert. Beispiele dafür sind GPS Messungen von Mobilgeräten, oder auch zeitlich hochauflösende Stromverbrauchsdaten gemessen von Intelligenten Stromzählern. Im Allgemeinen gibt es zwei sich widersprechende Anforderungen an die Daten: Zum einen ist es von hohem allgemeinen Interesse die Daten zu veröffentlichen um gesellschaftliche Ziele zu erreichen, dem gegenüber steht der Schutz der Privatheit einzelner.

Zeitreihen persönlicher Daten haben die Eigenschaft, das sie sowohl Informationen enthalten aber auch selbst identifizierend sind. Ein einfaches Entfernen der direkten Identifikatoren wie zum Beispiel des Namens oder der Adresse ist nicht ausreichend um die Privatheit der betroffenen Personen zu schützen. Mit Hilfe von Hintergrundwissen können Datensätze re-identifiziert, d.h. trotz fehlender Identifikatoren einzelnen Personen zugeordnet werden. Folglich müssen die bedeutungstragenden Daten selbst modifiziert werden um einen ausreichenden Schutz der Privatheit zu bieten. Im Folgenden werden Verfahren, die die Daten modifizieren 'Methoden zum Schutz der Privatheit'.

Abhängig vom Datenbestand können sich die Privatheitsanforderungen einzelner stark unterscheiden. Die Anwendung einer allgemeinen Methode zum Schutz der Privatheit hat den Nachteil, dass nicht alle persönlichen Informationen entfernt oder verfremdet werden.

Nach der Modifikation der Daten durch eine Methode zum Schutz der Privatheit, muss untersucht werden, welchen Einfluss die Modifikationen auf die resultierende Datenqualität hatten. Nur eine hinreichende Datenqualität stellt sicher, das die Anwendungen die auf den Daten basieren noch möglich sind und verlässliche Ergebnisse liefern. Abstrakten Maße, wie zum Beispiel der L1-Norm, hat den Nachteil, das der tatsächlich Einfluss der Modifikationen auf Anwendungen nur ungenügend reflektiert wird. Beispiel: Die Anwendung berechnet den Durchschnitt der Messwerte über ein bestimmtes Zeitintervall. Absolute Distanzen der Messwerte sind genau dann nicht von Relevanz, wenn der Durchschnitt trotz der Modifikationen durch die Methode zum Schutz der Privatheit erhalten bleibt. Somit geben abstrakte Abstandsmaße nur unzureichend Auskunft über die resultierende Datenqualität.

In dieser Arbeit werden Anwendungsspezifische Methoden zum Schutz der Privatheit und Spezifische Maße zur Bewertung der Datenqualität entwickelt und untersucht. Im Speziellen, werden sowohl Methoden als auch Maße im Kontext des sehr wichtigen und populären Szenarios des Stromnetz der Zukunft ('Smart Grid') untersucht.

Die Einführung des Smart Grid führt zu dem bereits vorhin erklärten Widerspruch von berechtigten Interessen: Es ist von gesellschaftlichem Interesse das Stromnetz zu erneuern und eine zuverlässige und umweltverträgliche Stromversorgung in der Zukunft sicherzustellen. Darunter zählt neben der Reduktion der CO_2 Emissionen auch die Einführung von neuen Anwendungen beispielsweise das intelligente Laden einer großen Zahl von Elektroautos. Eine Vielzahl von Gesetzen treibt die Einführung voran: Darunter fällt auch die Einführung von Intelligenten Stromzählern (sogenannten 'Smart Meter') in Privathaushalten. Sie erlauben die bi-direktionale Kommunikation und die automatische Übertragung von zeitlich hochauflösend gemessenen Stromverbrauchsdaten. Der Zugriff auf diese Daten ist nützlich für die erwähnten Anwendungen. Diese Daten fallen jedoch unter das Datenschutzgesetz, da sie persönliche Informationen enthalten. Privathaushalte haben somit ein berechtigtes Interesse daran, ihre Daten vor dem Zugriff Dritter zu schützen. Dieses Interesse steht im Widerspruch zum allgemeinen gesellschaftlichem Interesse der Weiterentwicklung des Stromnetzes.

Als allgemeingültiger Ansatz wird angenommen, das die Stromverbrauchsdaten mittels Veröffentlichung Dritten zur Verfügung gestellt werden. In diesem Fall ist keine vertrauenswürdige Instanz notwendig die den Datenzugriff regelt und der Zugriff der Anwendungen ist uneingeschränkt. Wie schon erläutert, werden die Daten zum Schutz der Privatheit verändert. Als Extremfall wäre vorstellbar alle Stromverbrauchswerte durch '0 kWh' zu ersetzen, jedoch schränkt das natürlich die Benutzbarkeit der Daten für Anwendungen stark ein.

Um beiden Anforderungen, also sowohl dem Schutz der Privatheit als auch das berechtigte Interesse Dritter für den Datenzugriff gerecht zu werden wird folgendes benötigt: Eine Methode zum Schutz der Privatheit deren Ziel es ist, private Informationen vor der Veröffentlichung zu entfernen. Zusätzlich, um die Nützlichkeit der Daten für Dritte bewerten zu können, ist ein Anwendungsspezifisches Maß für die veränderte Daten notwendig.

Die in dieser Arbeit vorgestellte Methode zum Schutz der Privatheit erlaubt Einzelnen die Spezifikation von Privatheitsbedürfnissen, sogenannten Geheimnissen. Unter bestimmten Voraussetzungen wird die Entfernung dieser Geheimnisse auch beweisbar garantiert. Geheimnisse betreffen meist bestimmte Gerätenutzungen die Rückschlüsse auf persönliche Verhältnisse zulassen: Ein Durchlauferhitzer der in Betrieb ist, erlaubt zum Beispiel den Schluss, das ein Bewohner des Haushaltes gerade duscht. Typischerweise hat ein sich in Betrieb befindender Durchlauferhitzer Auswirkungen auf eine bestimmte Menge von Messwerten in einer Zeitreihe. Um trotzdem systematisch und beweisbar die Privatheit zu schützen ist zunächst eine Transformation der Zeitreihe in eine abstrakte Darstellung notwendig. Dadurch ist es möglich, einzelne Informationen die 'versteckt' werden sollen zu isolieren und entsprechend zu verfremden. Um die allgemeine Anwendbarkeit auf typische Geheimnisse zu zeigen, wird die Methode mit Hilfe einer Menge von Geheimnissen die durch zwei populäre Anwendungen zur Informationsexktraktion aus Stromverbrauchsdaten definiert wer-

den getestet. Die Methode ist auch in anderen Domänen nutzbar, wenn gewisse Vorausetzungen erfüllt werden.

Das Anwendungsspezifische Maß erlaubt die Quantifizierung, zu welchem Grad die durch die Method zum Schutz der Privatheit veränderten Daten noch nutzbar sind. Genauer gesagt wird ein lokaler Energiemarkt implementiert. Hier wird Energie auf Basis von Zeitreihen von Verbrauchsdaten gehandelt. Vergleicht man das Marktresultat bei Benutzung von nicht-modifizierten Daten mit dem Resultat bei modifizierten Daten, erhält man ein Maß über die Auswirkungen der Veränderung. Theoretische Ergebnisse zeigen, das der Marktmechanismus wichtige Eigenschaften unter bestimmten Voraussetzungen wie die Anreizkompatibiltät, trotz systematischer Modifikation der Daten erhält. Die Veränderung der Daten hat Einfluss auf die Erlöse bzw. den Erfolg am Markt einzelner und unterschiedliche Methoden haben naturgemäß auch unterschiedliche Auswirkungen. Das vorgestellte Anwendungsspezifische Maß erlaubt den Vergleich der Methoden anhand von Real-Welt Daten. Auch wenn das vorgestellte Maß Anwendungsspezifisch ist, erlauben die Ergebnisse Rückschlüsse auf die Nutzbarkeit von Daten die durch Methoden zum Schutz der Privatheit modifiziert wurden im Allgemeinen: Das Maß zeigt, das sowohl Privatheit als auch Nutzbarkeit der Daten, unter bestimmten Voraussetzungen, erfüllbar ist. Da verschiedene Parameter des Energiemarkts und der Marktteilnehmer anpassbar sind, können allgemeine Rückschlüsse gezogen werden. Beispiele dafür sind die einstellbare Nachfrageflexibilität auf Haushaltsseite, die Abweichungen zwischen original und modifizierter Zeitreihe unterschiedlich in das Maß einfließen lässt.

Speziell für das Stromnetz der Zukunft zeigen die Ergebnisse das Methoden zum Schutz der Privatheit in einem lokalen Energiemarkt angewendet werden können, mit akzeptablen Auswirkungen auf die Anwendungen. Der Markt ist nach wie vor aus wirtschaftlicher Sicht effizient und verringert den CO_2 Ausstoß selbst wenn strenge Anforderungen an die Privatheit gelten und somit starke Modifikation der Daten notwendig ist.

Die Ergebnisse dieser Arbeit zeigen, das mit Hilfe von Methoden zum Schutz der Privatheit die widersprüchlichen Interessen zur Veröffentlichung der Daten auf der einen Seite und zum Schutz der Privatheit auf der anderen Seite vereint werden können. Da Datenbestände aus Zeitreihen von Stromverbrauchsdaten hinreichend komplex bezüglich dem Informationsgehalt und der Vielzahl der möglichen Anwendungen sind, liegt der Schluss nahe, das die gewonnenen Erkenntnisse auch in anderen Domänen gelten.

Contents

List of Figures

Chapter 1

Introduction

In the last few years, privacy concerns of individuals has become more important. More detailed and rigorous laws have been passed and the general awareness of privacy has risen in the society [15, 33, 36, 78]. With a vast amount of new technologies, personal sensitive information in time series is recorded especially more frequently. For instance, the time series of GPS data and the time series of energy-consumption data are measured by smart meters. Such data of individuals are subject to privacy legislation if they can be assigned to a single individual with little effort [15, 33]. If this is the case, such data sets require special treatment: Arbitrary access to the data is forbidden unless individuals accept disclosure. Judging from the general rising awareness for privacy, individuals will keep their data as private as possible. Additional effort is necessary to ensure security and limited access to stored data. However, such kinds of data are necessary to achieve important goals for the benefit of society. General speaking, the availability of data is beneficial to foster innovations [73]. This leads to two competing factors: First, the common interest of publishing a data set containing personal data for the general benefit of society, and second, the protection of individual privacy needs.

Although privacy and common access to important data for innovations is a general problem for time series of personal data, we explain our concepts in the context of one popular and important example scenario, the ‘electricity network of the future’ also known as the *smart grid* [50]. Modernizing the electricity-providing infrastructure toward the smart grid is a major concern for reducing CO_2 emissions and guaranteeing the security of supply at affordable prices. However, the modernization involves the collection of a huge amount of personal data, which is an inhibiting factor for the smart grid. In this scenario, two legitimate interests compete: The demand for establishing the smart grid and the privacy concerns. In this chapter, we will see that both are required by law. The concept for a smart grid to reduce emissions and guarante the security of supply is of national interest. In contrast, privacy laws restrict access to temporal fine-grain consumption data that smart meters collect. Further, we will elaborate on the contributions of this work, with the overall goal to fulfill both, privacy for individuals and access to data for the benefit of society.

In this chapter, we discuss the motivation of the smart grid (Section 1.1), including technical details about the smart meter (Section 1.1.2) and the legislation (Section 1.2). In addition to the technical details, the discussion also includes privacy concerns (Section 1.2.2 and 1.1.4). We conclude this chapter with an elaboration of the contributions in this work (1.3).

In the following, most of the explanations will be about the electricity network in Europe or in Germany. However, similar developments regarding liberalization, legislation, system architecture, and introduction of the smart grid are applicable to almost all countries that are striving to modernize the electrical grid.

1.1 Smart Grid: The Electricity Network of the Future

Despite some minor improvements, the current electricity network has not changed since the Industrial Age: Electrical energy is produced at central, large-scale power plants and distributed through a high-voltage system over large distances. The whole system is controlled centrally and designed to deliver high-quality electrical power to all consumers, even during peak times [49, 55]. The components of the system are mostly isolated and rather static: Despite the centralized source, it is difficult to inject energy. Consumers participate only passively, receiving only a monthly or yearly bill as feedback. Moreover, real-time monitoring is only limited to generation and transmission [41]. However, the existing network cannot answer the challenges of recent developments [31] concerning primarily energy prices, CO_2 emissions, and a rising demand. Prices for fossil fuels, e.g., coal and oil, keep rising [104]. Increasing energy costs conflict with the national interest of keeping the economy globally competitive [72]. Using fossil energy sources leads to a huge amount of CO_2 emissions, e.g., 87% of U.S. coal production is used for electricity generation [55] and therefore is a major contributor to greenhouse gases. Statistics show, that the electric power generation accounts for approximately 40% of the human-caused emissions of CO_2 [72]. Rising CO_2 emissions endanger keeping the climate targets [31], e.g., European road maps aim to reduce CO_2 emissions by 80% by 2050 [32].

Forecasts expect a rising energy demand in the next decades [93]. Furthermore, the electrification of the vehicle fleet will lead to additional load for the electrical grid [80]. Without any general change in the current grid and the generation of electricity, this will lead to higher prices and additional greenhouse gas emissions.

The smart grid is here, to challenge these issues. ‘It can be defined as an electric system, that uses information, two-way, cyber-secure communication technologies, and computational intelligence in an integrated fashion across the entire spectrum of the energy system from the generation to the endpoints of consumption of the electricity’ [41]. In particular, among other measures, this points to the installation of smart meters at the house connection. Smart meters allow (almost) real-time meter readings and establish a bi-directional communication channel between the energy supplier and consumer. For technical details of a smart meter, see Section 1.1.2. Using information and communication technology in the electrical grid is advantageous when dealing with fluctuating renewable energies, as explained subsequently.

Renewable sources like wind or solar energy reduce CO_2 emissions and cost for the energy production. However, these sources are different compared with a traditional power plant and difficult to coordinate. First, they are volatile by nature and the amount of energy produced cannot be controlled; it is only predictable. Second, they are usually distributed, e.g., solar panels on roofs of several private households, and therefore not centrally controllable. Communication technology enables real-time monitoring of supply and demand and allows for planning and distribution of renewable energy. This reduces the CO_2 footprint and the production costs [71].

Demand-side management is only possible if there is a communication channel to consumers. For example, [80] states that the current power grid infrastructure has spare capacity to support the penetration levels of hybrid vehicles ranging from 30 to 70% if they are charged during off-peak times. [81, 98] show that demand-side management is a powerful instrument in reducing peak loads and CO_2 emissions.

Using the smart-meter infrastructure allows the creation of dynamic tariffs. In the current grid, electricity costs a consumer mostly the same price per kWh throughout the whole day. Exceptions are only special tariffs with cheaper off-peak electricity and special contracts with industrial consumers. With smart meters, energy consumption is measured in real time and thus can also be accounted for differently for any time a day [111]. Dynamic tariffs, with high prices during peak hours, will encourage consumers to shift electrical load to off-peak hours and help to reduce emissions [38, 46].

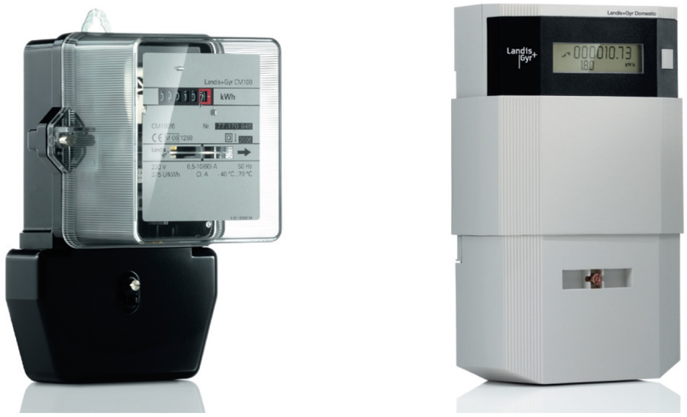
1.1.1 Architecture of the Electrical Grid

The electrical grid is usually divided into three different parts: *generation*, *transmission network*, and *distribution network.* [4] The generation includes all entities that produce power, which is transported over large distances through a high-voltage transmission network. The distribution network delivers the electric power from the transmission network to the consumer and is a medium/low-voltage network.

In this system we distinguish among the following involved parties. The *utility company* is responsible for power generation and insertion into the transmission network. We call the parties entrusted with transmission and distribution the *transmission system operator* and the *distribution system operator.* Additionally, the *metering service provider* measures the consumed energy and is responsible for the accounting at the consumer level. The *consumers* actually consume the transported electricity.

1.1.2 Smart Meter

A smart meter (see Fig. 1.1b) is typically installed at the house connection and replaces the 'Ferraris meter' (see Fig. 1.1a). The Ferraris meter, named after the Italian physicist Galileo Ferraris, is only capable of measuring the aggregate electricity consumption in an electromechanical way



(a) Ferraris meter [67]

(b) Smart meter [66]

Figure 1.1: Comparison of electrical meters

and is read manually, such as once a year. In contrast, a smart meter is considered to be ‘smart’ for reasons [66, 111] explained in the following. The core components of a smart meter are an electronic measurement device that includes a processing unit and a communication device. The electronic measurement enables (almost) real-time monitoring of the consumed energy in a household, and integrates the power-producing sites like photovoltaic panels. Communication devices may differ regarding the available networks, e.g., GSM, powerline communication, or integration into the households’ internet connection. The communication device enables the smart meter to automatically transfer the measured consumption data to the metering service provider, and to receive external information and commands. This includes the limitation of throughput resulting from the shortage of supply and demand-side management like starting to charge an electric vehicle. Most smart meters also have the capability to integrate other meters, e.g., water and gas, and forward the measurements through the communication device.

1.1.3 Current State of Development

In general, the smart grid is currently in the state that necessary preconditions are being established. There is a growing smart meter infrastructure that is necessary to provide most of the envisioned smart grid functionality.

For instance, in Germany, estimations predict that 30% of the total meters are smart meters until the year 2016 [13], whereas in 2009 the percentage was approximately 5% [85]. In Europe, the deployment of smart meters is very heterogeneous, e.g., in Italy and Spain, 95 – 100% of the installed meters have already been smart meters since 2009.

As a parallel development, pilot projects implementing specific parts of the smart grid ideas have already started. For instance, in Germany there is the E-Energy project[1] that includes six model regions. The government funds this project and there are a total of 140 Mio.€ available for the model regions. The regions follow different strategic orientations within a common set of goals for the whole project: In addition to energy efficiency, e-mobility, and the integration of renewable energies, these goals include IT architecture, security, and privacy.

The SmartGridCity is a technology pilot in the United States[2], specifically in Boulder, Colorado. The project deployed more than 23,000 smart meters. The main goals of this project are to increase efficiency and to develop a plan for a large-scale rollout.

In summary, the smart grid rollout is still in the beginning stages; however, the fraction of smart meters and thus smart grid–enabled technology will grow quickly in the next decade.

1.1.4 Technical Perspective of Privacy Concerns Regarding the Smart Grid

From a technical perspective, highly aggregated power consumption data, known from the traditional Ferraris meter and collected only once a year, does not indicate much about individual behavior. However, some information, like the number of people living in the household or whether a household uses inefficient devices, may still be observable. Power consumption data collected in short time intervals by smart meters contain a large variety of indicators for personal habits [78]. We discuss in greater detail which and how information can be extracted from such data in Chapter 2. [78] states that individuals might underestimate the possible privacy threats of their smart meter data and may even use opt-in services like the Google Power Meter[3], which helps to save energy but requires fine-grain data. Similar risks are that it is difficult to distinguish between sensitive and nonsensitive information in smart-meter data. Parts of the data may not be deemed sensitive and then disclosed [95]. For instance, legislation for the previously mentioned technical pilot ‘SmartGridCity’ (see Section 1.1.3) is a patchwork of different laws that may allow disclosure to an nonspecific degree for billing and load reporting [95].

Currently, the overall penetration of smart meters is relatively low, but it is possible that unsolved privacy problems may stall the smart grid [73]. One possible measure to increase privacy protection would be to explicitly define which kind of information is really required for the performance of applications like system balancing, demand reduction, and distribution network operation and planning [79]. In this thesis, we introduce a method for removing specific private information in such data. Additionally, we introduce an application-specific measure—the local energy market —and quantify the negative influence of privacy-enhancing methods on the market. Those contributions are further explained in Section 1.3. From a technical perspective, the interest of a large-scale rollout competes with the privacy issues in the smart grid. In the next section, we discuss the juridical perspective, including the legislation for establishment of the smart grid, as well as the privacy legislation.

[1] http://www.e-energy.de/

[2] http://smartgridcity.xcelenergy.com/

[3] http://www.google.com/powermeter/about/

1.2 Legislation

Legislation related to the smart grid that are relevant for this work basically cover two aspects: First, because the introduction of the smart grid is a goal of national interest, several laws and edicts have been introduced to foster its development (Section 1.2.1). Second, the data collected by smart meters are subject to privacy legislation (Section 1.2.2).

1.2.1 Legislation Fostering Smart Grid Development

The first step in establishing a smart grid is the liberalization of the electricity market. This allows different parties to access the grid and fosters competition. In the European Union, the Directive 96/92/EG [34] claimed the detachment of electricity companies to multiple division according to the grid architecture (see Section 1.1.1) in production, transmission, and distribution. In Germany, this directive led to the Renewable Energy Act (EnWG [26]) in 1998. Modifications to this law in the following years introduced even more liberalization, e.g., the liberalization of the measuring in 2008. This includes that law requiring the installation of smart meters for newly constructed or majorly refurbished buildings at the beginning of 2010. Energy suppliers have to offer tariffs, depending on the load or daily times to consumers.

In addition to fostering liberalization, the EU brings the development of the smart grid forward with research road maps and initiatives: The European Strategic Energy Technology Plan (SET-Plan [30]) is an industrial initiative to increase the fraction of renewables in the electricity grid up to a completely decarbonized electricity production in 2050. The European Technology (ETP) SmartGrids was set up in 2005 and created a research agenda for 2020 [31]; the goals were renewed in 2012 [32]. The documents contain research topics and priorities to reduce emissions and guarantee the security of supply in the near future (2020 and 2035). These initiatives foster the development and the research of smart grid technology in the European Union.

In the United States, the policy in the 42 United States code ch. 152, sub chapter IX §17381 is relevant for the smart grid legislation. It establishes a Smart Grid Advisory Committee and a Smart Grid Task Force. Their goal is to develop smart grid technologies, define standards, and plan a transition of the existing U.S. electricity grid to a smart grid.

Legislation fosters the development and deployment of smart grid technologies in Europe and as in the United States. Although measures may differ, the goals are the same: Communications technology should enhance the integration of renewable energy sources for a reduction of emissions and for guaranteeing the security of supply.

1.2.2 General Privacy Legislation

Privacy is a fundamental right of natural persons. It is motivated by the so-called freedom ‘to contribute to economic and social progress, trade expansion and the well-being of individuals’ [33]. The right of privacy is not only covered by European legislation; because [33] is a directive, it has also found its way to national legislation of the EU members, such as the German Privacy Act [15]. There is similar jurisdiction in the United States, such as the 19th Annotation to the First Amendment [110].

For the sake of simplicity, we explain the German Privacy Act [15], which carries out the EU directive [33], in greater detail and omit details of other state-specific legislation. Privacy legislation in different states is mostly distinguishable in the integration in the law system. For instance, privacy legislation in the United States is an annotation of the first amendment, whereas laws in the EU fulfill directives.

Data are personal and subject to privacy legislation if they reflect the living conditions of individuals [15]. This includes wealth, working hours, and leisure activities. Additionally, information like the number of people present in a household may also be personal if it reflects the relationships of an individual. Furthermore, the information has to concern an identified or an identifiable person [15, 33]; identified information usually contains names or addresses. A person is considered as identifiable if data can be linked to an individual with reasonable means, e.g., combinations of zip code, age, and sex may be enough to re-identify a person without having any available direct identifier. Processing and working with personal data have the following consequences: Access to and use of personal data have to be legitimated. The individual needs to confirm whether his or her data can be used for a specific purpose. This permission only holds for that purpose and only for the data required to process this specific purpose, e.g., if aggregated data can be used, the access to fine-grain data is not allowed [15]. These principles are called data minimization and data avoidance. The so-called direct survey principle is also important for the smart grid scenario: This principle basically regulates by law that personal data can only be collected if the individuals participate directly in this process.

Privacy Legislation in the Smart Grid

The consumers (households), the utility company, and the metering service provider are the important parties when considering privacy laws in the smart grid. The privacy of the consumers is protected, the metering service provider collects the private data, and the utility company needs the data for accounting. In addition to the German Privacy Act, laws relevant for the individual's privacy are the Renewable Energy Act (EnWG [17]) and the metering access ordinance (MessZV [26]). These laws cover the installation of smart meters in private households and implement the terms introduced in the German Privacy Act. Because of the principle of direct survey, the metering service providers have to explicitly ask the consuming private households for permission to automatically access the data from the smart meters. The principles of data minimization and avoidance require that the smart-meter data that are forwarded to the utility company have to be aggregated according to the tariffs. For instance, if a private household has a tariff with prices changing every hour, the utility company is only allowed to get access to hourly aggregated data from the metering service provider [17, 26, 63]. Furthermore, according to the German Privacy Act, the data can only be used for the purpose of accounting.

1.3 Contributions

We evaluate the proposed methods and measures within the smart grid scenario. As already explained, the modernization of the existing electricity grid is of common interest to fulfill CO_2 goals, e.g., by integrating renewable energy or by introducing intelligent charging for a large number of electrical vehicles [32]. Thus, there are a number of reasons to promote the expansion of the smart grid. A central role of the smart grid is the introduction of the smart meters. Smart meters collect the time series of electricity consumption from private households in short time intervals [66]. Such data are necessary for further development of the smart grid, but contain private and sensitive information [83]. Smart meters enable two-way communication and transfer of consumption and production data among all participating parties, including private households. Special treatment of such data is necessary if it can be assigned to individuals with little effort. Personal identifiers such as names or addresses directly link data to individuals. Such identifiers can be replaced with pseudonyms. However, we show that it is still possible to assign data without personal identifiers with minimal effort to individuals. This process is called re-identification and is our first contribution.

C.1 Method and features to re-identify the time series of smart-meter data: We systematically analyze the identifying degree of features for the time series of smart-meter data. Results show that, depending on the data set, the identifying degree of features differs. Furthermore, we propose a method that allows the re-identification of households to time series with little effort (Chapter 2).

The effort in computational time is in many cases below 1 min on current standard hardware. We also see that the probability re-identification depends on the data size considered. The more time series that are involved, the more likely are similar features of consumption records. This reduces the re-identification rate. However, we show that re-identifying households by their smart-meter data is possible and by an order of magnitude better than random guessing. Consequently, it is possible to reference the time series of smart-meter data to individuals, making such a data set subject to data-protection legislation. The proposed features are possible information that an individual wants to remove before disclosure to hinder re-identification [11].

Legitimate privacy interests compete with the common interest of introducing a modernized electrical grid. The core of privacy legislation is the so-called right to informational self-determination of each individual: Each individual has the right to decide who and for which purpose one has access to certain personal information [15]. Numerous approaches exist for extracting different information from the time series of smart-meter data. In addition to the features proposed in the re-identification method, smart-meter data also reflect the presence of individuals, which appliances are active, and in which state (i.e., which TV program the household watches) [43]. Privacy-enhancing methods generally modify a given data set with the objective to preserve the privacy in a data publishing case. In particular, for smart-meter data, this requires removing information that individuals deem sensitive. In turn, with the help of such a method, individuals determine which information is provided for applications leading to the benefit of society. We contribute PACTS, a privacy-enhancing method for the time series of smart-meter data.

C.2 PACTS provable privacy method: We contribute a privacy-enhancing method for a time series that allows each individual to define the privacy requirements. Each time series is handled and published in isolation, as privacy requirements may differ. The method provides provable guarantees for the removal of the specified information.

State-of-the-art approaches for providing provable guarantees usually focus on relational or aggregate data [5, 29], whereas PACTS focuses on the privacy enhancement of individual time series. Challenges for such a method is that the information deemed sensitive usually cannot be referenced to a single time-value pair, but is a sequence or set of values. PACTS proposes a general method for using abstracted representations to isolate private information. PACTS is capable of hiding private information that is required by the proposed re-identification method to reference time series to households and the extraction of appliance states by a recent non-intrusive-appliance-load monitoring approach [8].

Privacy-enhancing methods, including PACTS, modify the provided data set. It is questionable whether the modified data are still useful for applications. An extreme example is that all the electricity consumption values are set to zero at each point of time. Although this will obviously preserve privacy, the data quality is low and the applications cannot provide any benefit for society. Thus, we require a measure for determining the influence of privacy-enhancing methods on data quality. In particular, for the time series there are abstract measures to reflect the distance between the original and the privacy-enhanced one [115]. However, abstract measures do not reflect the actual effect on applications for a specific scenario. For example, the L1 norm reflects absolute distances, but does not judge the data quality for an application requiring correct averages. Additionally, resulting numbers give a rather abstract idea of the actual impact. In contrast, we develop an application-specific measure that is meaningful in the smart grid scenario and returns intuitive understandable measures. This allows us to gain insight into the effects of privacy methods and their parameters and allows for the comparison of different approaches.

C.3 Local energy market: Application-specific data-quality measure: We integrate privacy-enhancing methods in a local energy market. The comparison of market outcomes with privacy-enhanced and unmodified data results in intuitively understandable measures such as welfare and CO_2 emissions.

Measuring the data quality with the help of local energy markets is challenging: It requires a design of local market, that is aware of privacy-enhancing methods and includes the integration in the remaining power grid. Simulations require models for supply and demand, including prices and valuations of participants.

The local energy market depends on the smart-meter time series provided by individuals. The market outcome is useful as an intuitive measure for the impact of privacy-enhancement methods. In the scenario itself, antagonism exists between privacy and benefit to society. Local energy markets are a powerful tool for the automatic distribution of renewable generated electricity. However, they are a threat to participating individuals because they require publication of the electricity demand. Modifying the time series to protect the privacy decreases allocation efficiency and influences the theoretical properties of such markets. To understand the influence of privacy-enhancing methods, we investigate the interplay in detail in our fourth contribution [12, 57].

C.4 Impact of privacy enhancement on electronic markets: We evaluate the theoretical and numerical effect of privacy-enhancing methods on electronic markets by example of the local energy market. We will prove theoretically that the market mechanism keeps important properties like incentive compatibility. The conducted numerical evaluation shows that the overall impact on measures like welfare is low and controllable, and that storage systems are capable of mitigating the negative effects of privacy enhancement.

In addition to PACTS, several other privacy-enhancing methods exist. To derive theoretical results, we need to define properties that are general enough to cover a number of such methods and specific enough to have a meaning for these markets. Numerical evaluation requires real-world data to implement models of supply and demand. The investigated scenario is integrated into the smart grid context; however, the results are applicable to time series in general: Utility and privacy enhancement do not exclude each other. Applying an application-specific measure allows the comparison of enhancement methods. The results hold for other applications as well, if the effect of changes in the time series have comparable consequences for the utility. For instance, increasing or decreasing time series values may have a different impact. Furthermore, the local energy-market measure is adaptable to various configurations, e.g., if there is undersupply, decreasing time series values may have no impact on the revenue at all.

For the smart grid scenario itself, the results clearly show that privacy-enhancement methods are applicable to a local energy market with acceptable consequences: The market is still efficient in economic terms and reduces CO_2 emissions even with strict privacy requirements.

1.3.1 General Applicability of Contributions

Within the smart grid scenario we evaluate the proposed methods and measures. However, the contributions are general applicable in the context of privacy for time series of personal information: The proposed re-identification method for time series (**C.1**) is independent of the features considered. Thus, it is very likely that such features can also be found for a different data set consisting of time series of personal information, e.g., GPS trajectories. The same considerations can be applied to contribution **C.2**. Information deemed sensitive might be distributed amongst several points of time. An abstracted representation can be used to isolate certain information and to provide provable guarantees according to PACTS. Contributions **C.3** and **C.4** show in particular, that it is possible to ensure privacy of individuals and provide useful data for applications in a real-world scenario. The results indicate, that this is also possible for applications dependent on different time series and might even be applicable to privacy methods for any personal data. Respecting the privacy of individuals will lead to a lot more data and information publicly available fostering research as well as applications for the benefit of society in many fields.

Chapter 2

Re-identification and Privacy Threats

Various devices record and store the time series of individuals. Popular examples are GPS devices, smart meters, or self-tracking devices that measure the body functions. Data that are subject to privacy legislation require special treatment: Access is limited to contractually established, predefined purposes and parties. Data fulfilling the following two properties (see Section 1.2.2) make it subject to privacy legislation:

1. The data have to be personal in that it reflects the living conditions of individuals.

2. The information has to concern an identified person or is identifiable with minimal effort.

Smart-meter data reflect personal details ('living conditions') because information about habits and running appliances can be extracted. This is not straightforward and requires the help of complex models, such as appliance load signatures. However, the extraction of such information is possible and we explain the approaches from related works in Section 2.1.1. Figure 2.1 illustrates the information extracted from a time series of smart-meter data.

To be relevant for privacy legislation, data also have to concern an identified or identifiable person. Identifiable means that it is possible to reference a data set to an individual. Usually, this is straightforward under the presence of identifiers like the name or the address. We will show that the reference to individuals is possible even without the presence of such identifiers. Consequently, the time-series data itself is identifying. In Section 2.2 we show how re-identification can be achieved, with the help of external knowledge concerning features of the consumption, and is illustrated in the following example:

Example 1 ***(Re-identification example)*****:** A network operator works with energy-consumption data from a certain area. Because the operator is not involved in billing or cashing, it does not know the identity of the households from where the data are extracted. However, an employee of the operator knows that his or her neighbor typically uses the coffee machine at 7.15 a.m. and the microwave at 1.30 p.m.. Suppose that only one time series from the consumption data has

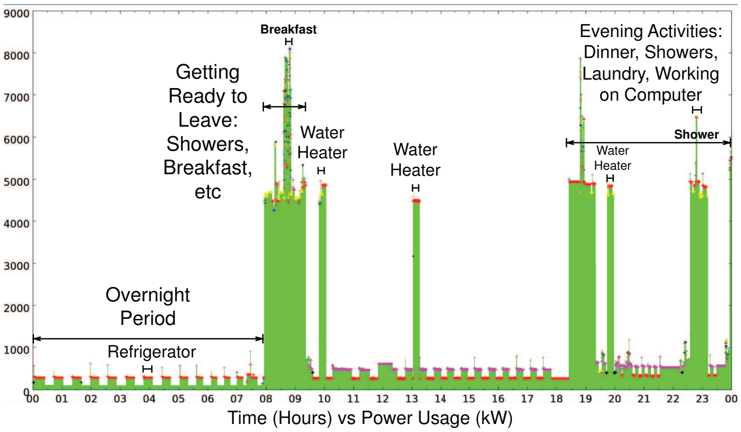


Figure 2.1: Example smart-meter data with extracted information on activities [83]

these characteristics. In this case, the employee could find out which consumption data belong to his or her neighbor, and explore the neighbor's entire consumption history. □

This example illustrates that the re-identification and information extraction are orthogonal to each other. Even if it is impossible to re-identify smart-meter data, it could still be possible to extract sensitive information, and vice versa. Consider a set of two identical time series: It is not possible to compute pseudonyms that distinguish both time series, but information such as the daily routines could be extracted.

In this chapter, we analyze to what extent anonymous energy-consumption records are prone to re-identification and fulfill contribution **C.1**. In particular, we are interested in the effectiveness of simple statistical measures to this end. Furthermore, we investigate which features of the energy consumption of the households are particularly well-suited to re-identify consumption data. Our findings indicate, that privacy obligations apply to smart-meter data stripped from personal identifiers. Our study is based on the observation that almost all daily activities, from making breakfast to relaxing with a game console, influence the energy consumption. Because the daily routine is influenced by many aspects of the household, e.g., employment status, hobbies, or the number of persons, features of the energy-consumption data should be inherently identifying for many households. Furthermore, we consider features like the aggregated consumption per day or the time of the first peak demand in the morning, and we analyze to what extent we can use these features for re-identification. Instead of striving for complete sets of features or sophisticated algorithms, we are interested in finding out whether straightforward features and simple statistical measures are sufficient for re-identification of consumption data. Our goal is to show that potential for misuse of smart-meter data is high. With simple measures, non-experts would be able to perform re-identification. Straightforward features that could be estimated or observed by anyone would increase the privacy threat even more.

In particular, for **C.1** we contribute the following:

1. We identify and analyze a number of energy-consumption features, and we quantify to what extent they can be used for re-identification;
2. We describe an analytical framework for the re-identification of energy-consumption data according to the consumption features; and
3. We measure to what extent it is possible to systematically re-identify households based on consumption features.

The remainder of this chapter is structured as follows: We begin with the discussion of related works covering information extraction and re-identification (Section 2.1) and continue with our re-identification approach (Section 2.2) before we conclude (Section 2.3).

2.1 Related Work

In general, we distinguish between information extraction and re-identification. Given a record of personal data, *information extraction* means to infer personal information from that data record. This term summarizes all kinds of information extraction, including probabilistic cases that extract information and a corresponding probability, i.e., the confidence that a certain device is actually running. *Re-identification*, in turn, refers to the process of re-assigning data without an identifier like a name or an address to an individual. Data that contain personal information and are referable to individuals are subject to privacy legislation.

2.1.1 Information Extraction

One popular class for extracting information from smart-meter data are the so-called non-intrusive appliance-load monitoring (NIALM) methods. The term was first introduced in [45], summarizing methods that take an aggregated power consumption value and then return information about running devices and their states. In contrast to *intrusive* load monitoring, it does not require placing measurement devices on individual appliances. As a first approach [45], trains finite state models and signatures of appliances as external knowledge. The recent INDiC approach [8] improves the original method [45]. Depending on the actual appliance, the state guessed up to 89% of the cases as being correct. Compared with other approaches, it is simple and it detects appliances accurately. In Chapter 3 we use INDiC as a generality test for the proposed privacy-enhancing method. Thus, we explain [8] in greater detail. INDiC assumes that each appliance has a number of states with different extents of power consumption, and an appliance can only be in one state at a time. In this case, disaggregation is a combinatorial optimization problem, namely, finding the optimal combination of appliances in different states while minimizing the error. INDiC requires a special data set for training. It consists of the total power consumption of the household, usually measured by a smart meter at the mains' connection, and the consumption of single appliances in question. The test data set for the NIALM approach only consists of the mains' time series. For example, the REDD data set (see Section 3.6) qualifies as training and a test data set because it contains the main power consumption in addition to appliances in isolation. INDiC conducts the following steps:

1. *Initialization/Preprocessing:* The temporal resolution of the power measured at the mains and the appliances may differ; so, in the first step, INDiC harmonizes the time series by, e.g., downsampling;

2. *Training phase:* INDiC conducts a clustering step of the appliance data to extract the typical states of an appliance. For example, a 'light bulp' has two states: It consumes no electricity when switched off and a fixed amount when switched on. Furthermore, the mapping of appliance states to the main power consumption is calibrated, because the household may have additional devices that are not monitored in isolation; and

3. *Combinatorial optimization on test set:* The last step is the actual NIALM step. INDiC solves a combinatorial optimization problem, in which the sum of all appliance state consumption values have to be equal to the total consumption at a specific point in time.

NIALM is also possible without the training step and is called *unsupervised disaggregation* [42, 61, 83]. For example, the methods proposed use Hidden Markov [61] models to detect appliances or to label activities in the household [83]. Numerous approaches exist; see [121] for an overview. Obviously, using appliances reflects living conditions and gives insight into the typical activities of the household.

These approaches can detect whether a household has the TV switched on. Additionally, [43] can extract which program is running. It is a well-known fact that current televisions require less power when displaying dark or black images compared with bright ones, because the backlight is dimmed in dark scenes. This leads to an identifiable signature of the current running program [43].

Thus, the time series of power consumption data as measured by smart meters is personal. Combined with further external knowledge, like the times of church services in communities nearby compared with the presence or absence in a household on Sundays, the smart-meter data provide insight into the private lives of individuals.

2.1.2 Re-identification

It is well-known that personal data, even without containing identifiers like the name or the address of a person, can be re-identified with the help of external knowledge. For example, a study from 1986 shows that 63% of U.S. citizens are identifiable by the combination of the date of birth, gender, and ZIP code [22]. A succeeding study from 1997 was conducted on the voting list for Cambridge, Massachusetts, and contained the demographics on 54,805 voters [107]. 97% of the individuals are identified by the full postal code and the birth date, and still 69% with only a birth date and a 5-digit ZIP code (the full ZIP code consists of 6 digits). In the following, we refer to attributes that do not contain identifiers but help in re-identifying records quasi-identifiers [108].

The construction of these quasi-identifiers give way to re-identifying records that contain personal information. In [108], external knowledge from a publicly available voter list was combined with a public data set of health records, leading to a re-identification of a governor of Massachusetts. Re-identification is also an issue for other data sets like the published AOL search records [7]: The published data set contained only the searched terms combined with a unique key that did not qualify as a personal identifier. With the help of a fraction of the searched terms, the adversary was able to infer the address of an individual and re-link all of the searched terms. The remaining terms that were also linked to that person gave insight into her personal life, e.g., ‘numb finger’, ‘60 single men’, and ‘dog that urinates on everything’.

[39] proposed a method for securely computing pseudonyms for smart meters without requiring a trusted third party. Such a party may facilitate re-identification by combining unique network addresses with pseudonyms. However, the method proposed in this work does not rely on any information in addition to consumption-related external knowledge.

The time series of GPS trajectories are known to be identifying [86]. In particular, regular locations in the morning usually refer to the home of an individual and the first route usually ends at the workplace. The time series of smart-meter data [53] introduce a way for re-identification with the help of two attack vectors using anomaly-detection-behavior pattern matching. [53] used a rather complex solution that is fixed to specific attacks. It is difficult to argue that these attacks still require only minimal effort. In contrast, the proposed method is a general one, applicable to a number of intuitive features and can be easily extended to new features.

2.2 Re-identification Approach

In this section, we introduce our re-identification approach [11]. The concept of re-identification has already been shown for several other data sets. We will show that the time series of smart-meter data can be re-identified by an adversary with a high probability with simple statistical measures. We extract different features of the energy consumption data, e.g., the first or the last peak of the day, and provide an analytical framework to quantify how well these features can identify households. If an adversary has external knowledge of a feature, a household can be re-identified. Some features are easily obtainable by others, e.g., the time of the first peak is related to the time that a household gets up and some others require more knowledge about the consumption over a specific period of time.

Depending on the assumed external knowledge, we are able to re-identify up to 80% of the tested households. High re-identification rates are possible within 10 s of computational time on current hardware. We will show, that adding complexity does not necessarily increase re-identification rates by much.

2.2.1 Common Notation

For the sake of consistency, we introduce a small set of common notations for definitions used in this chaper as well as throughout the whole dissertation. This covers the notation of time series, the relevant domains, and individuals to whom the time series belong.

Notation 1 (Time domain $\mathcal{T}$): The time domain $\mathcal{T}$ is a countable infinite set and contains all points of time t that defines a time series.

Notation 2 (Value domain $\mathcal{V}$): A time series assigns a value v to each point of time in $\mathcal{T}$; the set $\mathcal{V}$ contains all possible values v.

Notation 3 (Time series): A time series $f(t)$ maps points of time $t \in \mathcal{T}$ to values $v \in \mathcal{V}$.

Notation 4 (Individuals): In this work, time series are always data from individuals (respectively their households); each time series f represents the data of an individual.

Note: to ease our presentation, we summarize the term 'individual' as a single person and a small group of persons who are still relevant for privacy concerns, e.g., a household.

2.2.2 Study Methodology

This section provides the theoretical background of our study. To investigate the identifying degree of energy-consumption data, we analyze real-world smart-meter readings with different periods. We take a set of time series $\mathcal{F}$ with the time domain $\mathcal{T}$ measured by smart meters of $n = \|\mathcal{F}\|$ individual households $\mathcal{I}$. Our goal is to show that having certain external knowledge $\mathcal{K}$ is enough to identify an individual $p \in \mathcal{I}$. In the following, we refer to this data set as the 'test data'.

Distinguishing the energy consumption of households with extremely different characteristics tends to be rather simple: For instance, a single-person household has a significantly lower energy consumption than a multi-person household, or daytime employees will have different peaks than shift workers. Instead, we will show that it is still possible to re-identify households with similar energy-consumption developments. This challenging setup requires an analytical framework that uses a combination of consumption features for re-identification. Re-identification, while having precise knowledge of the actual data, is not challenging. Thus, we assume that an attacker has external knowledge $\mathcal{K}$ from any source, but not from the test data. We will show instead, that for most of situations, re-identification is possible with only imprecise information.

In our model, external knowledge $\mathcal{K}$ consists of several feature values that describe power consumption properties (see Definition 1), a tolerated error that determines possible imprecision (see Definition 3), and the importance of each feature to the others, which we subsequently call 'weight' (see Definition 2). Features from external knowledge may differ from the features of the test data even if both belong to the same individual. For example, the first peak in the morning may differ a few minutes each day. Taking this into account, the adversary also knows the tolerated error, which is the maximal difference that a feature for the same household for different periods of time is assumed to have. Each weight determines the identifying degree of a certain feature in comparison to the others in a data set. For example, if the adversary assumes that the total power consumption distinguishes the households better than the time of the first peak, the weight of the total power consumption will be higher.

Definition 1 (Feature ϕ): A feature is a calculation rule describing a certain, possibly identifying property of an individual's electricity consumption. The value of a feature ϕ_f refers to an actual result of the calculation for time series f. □

Definition 2 (Weight ω_ϕ): The weight ω_ϕ is a factor discriminating the importance of feature ϕ in relation to other features of a data set. □

Definition 3 (Tolerated Error δ_ϕ): The tolerated error δ_ϕ quantifies the maximum distance of two features to be considered from the same individual. □

Definition 4 (External Knowledge $\mathcal{K}$): The external knowledge of an adversary consists of a set of features Φ, a set of weights Ω, and a set of tolerated errors Δ.

$$\mathcal{K} = \{\Phi, \Omega, \Delta\}$$

□

To mimic an adversary having external knowledge, our approach includes a training phase. In this phase the re-identification framework learns the weights for each feature with the help of a training data set. The training data set is fully known. More specifically, our framework conducts the following steps:

1. Training:

   - *Feature and Distance Computation:* We divide the training data set into two distinct periods. Based on the features described in Section 2.2.2, we compute the feature values of each time series in the first and second period of the training data set. Additionally, we compute distances between the features in both periods. The calculated distances in combination with the feature values of the known households allow the computation of weights.
   - *Weights Computation:* Features may vary in their spread of values: A high spread is more likely to differentiate individual households than a low one. Thus, features should have different weights for the final re-identification. We use a static, linear optimization and an integer linear optimization approach in addition to determine weights. We investigate whether higher computational complexity increases the re-identification performance (see Section 2.2.2).

2. Re-identification:

   - The final step is to calculate the weighted distance between the precalculated features of a household in combination with the results of the training phase as external knowledge

	Absolute Difference	**Relative Difference**
Consumption	Overall Consumption Minimum Consumption 0.9-Quantile	Maximum Consumption Standard Deviation Frequency of Mode
Consumption During Time Interval	Consumption M–F, 4 a.m.–8 a.m. Weekend Consumption	Consumption M–F, 10 a.m.–4 p.m. Consumption M–F, 9 p.m.–2 a.m.
Time	Average Wakeup Hour Average Bedtime Hour	

Table 2.1: Electricity-consumption features

and an anonymous consumption record. A household is re-identified if the distance to the correct household is lower than the distance to any other one.

The rest of this section is structured as follows: First we introduce the features we use for re-identification and then describe the necessary algebraic framework.

Determining Features

Based on the insights of Section 2.1.1, we assume that the electricity consumption of a household reflects the daily routine. In the following, each defined feature covers a certain aspect of the routine. Features that are best suited for re-identification have the following properties: First, because we prefer features that tend to be identifying, a feature should not change as long as the household keeps the daily routine. For example, the 'average wakeup hour' will not change permanently unless the individuals of the household change their way of life significantly. Additionally, for different time periods of the same household, the value of the feature should stay the same, whereas it differs from values of the same feature for other households. Second, it is possible for an adversary to guess the value of the feature by observing the household's way of living. In particular, estimating the feature value does not necessarily require access to the electricity consumption data of the household in question.

In general, we assume that there are no limits regarding the external knowledge of an adversary. Thus, every possible feature could be contained in the adversary's knowledge. Still, the proposed features cover a large variety of different aspects, including intuitive, easy-to-determine external knowledge.

In the following, we define the features for re-identification. Features can take three different combinations of time-series properties into account: the *consumption* values solely, the *consumption during a time interval*, and the *time*. These three properties form a categorization of features. Table 2.1 summarizes the defined features. For the sake of exposition, we refer to $f \in \mathcal{F}$ as the time series and $\mathcal{T}$ as the time domain by which a feature is defined.

Consumption These features rely on the measured consumption values of individuals only.

Definition 5 (Overall Consumption (OC)): The overall consumption is the sum of all consumption values:

$$\phi_f^{OC} = \sum_{\forall t \in \mathcal{T}} f(t)$$

□

Definition 6 (Minimum Consumption (MinC)): The minimum consumption is the lowest of all consumption values in the time domain:

$$\phi_f^{MinC} = min\,(\forall t \in \mathcal{T} : f(t))$$

□

Definition 7 (Maximum Consumption (MaxC)): The maximum consumption is the highest of all consumption values in the time domain:

$$\phi_f^{MaxC} = max\,(\forall t \in \mathcal{T} : f(t))$$

□

Definition 8 (Standard Deviation (SD)): The calculation of the standard deviation requires the mean value of the time series. Let $N = \|\mathcal{T}\|$ be the number of time-series values, then the mean $\overline{f}$ is

$$\overline{f} = \frac{\sum_{\forall t \in \mathcal{T}} f(t)}{N}$$

The actual feature value is calculated as follows:

$$\phi_f^{SD} = \sqrt{\frac{1}{N} \sum_{\forall t \in \mathcal{T}} (f(t) - \overline{f})^2}$$

□

Definition 9 (0.9**-Quantile):** The quantile is a statistical measure representing a threshold that divides the ordered sequence of consumption values in predefined fractions. The 0.9-quantile $\phi_f^{0.9Q}$ is the threshold that divides the consumption values of f in the upper 10% and lower 90%. □

Definition 10 (Frequency of Mode (FOM)): To determine the frequency of mode, we calculate the number of each unique value $f(t)$. The frequency of mode ϕ_f^{FOM} is the number of occurrences of the most frequent consumption value, i.e., the highest value of the conducted calculation. □

Consumption during time interval: These features consider consumption and the point of time when the electricity was consumed.

Definition 11 (Consumption Mo-Fr h_1-h_2 (MF)): This feature is the sum of all consumption values during the hours h_1 and h_2 on a weekday, i.e., between Monday and Friday. For the sake of simplicity, we first filter the time domain and get the points of time in question. Let $isWeekday(t)$ return true if t is on a day of the week and false otherwise. Furthermore, let $hour(t)$ return the hour of a day represented by t:

$$\mathcal{T}' = \{t \in \mathcal{T} | isWeekday(t) \wedge h_1 \leq hour(t) \leq h_2\}$$

The feature value is the sum of all consumption values during the already filtered time span:

$$\phi_f^{MF(h_1,h_2)} = \sum_{\forall t \in \mathcal{T}'} f(t)$$

As features, we consider time intervals on weekdays between 4 a.m.–8 a.m., 10 a.m.–4 p.m., and 9 p.m.–2 a.m. □

Definition 12 (Weekend Consumption): Similar to the consumption during the weekdays, this feature represents the complete weekend consumption. Let $isWeekend(t)$ return true, if t is on a Saturday or a Sunday and false otherwise.

$$\mathcal{T}' = \{t \in \mathcal{T} | isWeekend(t)\}$$

The feature value is the sum of all the consumption values at $t \in \mathcal{T}'$:

$$\phi_f^{MF(h_1,h_2)} = \sum_{\forall t \in \mathcal{T}'} f(t)$$

□

Time These features determine points of time when energy consumption increases or decreases.

Definition 13 (Average Wakeup Hour (WH)): The average wakeup hour is the average time of day when the first significant increase in electricity consumption occurs. The calculation considers only weekdays, as wakeup times on weekends may vary. A single wakeup hour is h in two cases: First, if the consumption in hour h is 30 W higher than the consumption in the hour before; second, if the sum of consumption during the hour h and the hour before $h-1$ is 40 W higher than the consumption in the second to last hour before $(h-3)$. The feature $\phi_f^{WH} = \overline{h}$ is the average value of all calculated wakeup hours h. □

Definition 14 (Average Bedtime Hour (BH)): In turn, the average bedtime hour is the first significant decrease in power consumption in the afternoon after 4 p.m. The hour h is the bedtime hour if the energy consumption is 30 W lower than in the hour before. The feature $\phi_f^{BH} = \overline{h}$ is the average value of all calculated bedtime hours h. □

The list of features is obviously not complete; one can easily think of other features related to the lifestyles of individual households. However, the goal of this study is not to provide a complete list of possible features, but rather to show that re-identification is possible with simple means.

Algebraic Framework

Remembering that the goal of this study is to re-identify an individual's time series in a set of time series with any identifier. The algebraic framework precisely describes the way we calculate a score for each of the time series in question and the procedure of 'guessing' the right time series. The computed score is a weighted distance between the features calculated in a training period as simulated external knowledge and the features of a time series in question. We determine the weights in three different ways, allowing a comparison between the computational effort and the re-identification performance.

In the following, let $\mathcal{I}$ be the set of individuals $p \in \mathcal{I}$ in question. f_p denotes the time series that belongs to p. Furthermore, we distinguish between the set of time series in the training period $\mathcal{F}^\alpha$ and the re-identification period $\mathcal{F}^\beta$. The time series in both sets are defined as $\mathcal{T}^\alpha$ respectively $\mathcal{T}^\beta$, whereas $\mathcal{T}^\alpha \cap \mathcal{T}^\beta = \varnothing$ holds. For each time series $f_p^\alpha \in \mathcal{F}^\alpha$ we know the individual $p \in \mathcal{I}$. In turn, for each time series $f_{p'}^\beta \in \mathcal{F}^\beta$ we want to determine whether $f_{p'}^\beta$ is from the same individual, i.e., $p = p'$.

We consider absolute and relative distances of these features. Relative distances might be more informative than absolute ones, e.g., if the maximum consumption is 90 W, the relative distance to 120 W is as significant as the distance from 3000 W to 4000 W. The distances of feature ϕ for time series f_p^α and $f_{p'}^\beta$ are defined as follows:

Definition 15 (Absolute distance d_{abs}): We define the absolute distance between f_p and $f'_{p'}$ as a difference of the feature values ϕ:

$$d_{abs}^{\phi}(f_p^\alpha, f_{p'}^\beta) = \left| \phi_{f_p^\alpha} - \phi_{f_{p'}^\beta} \right|$$

□

Definition 16 (Relative distance d_{rel}): In turn, the relative distance between f_p and $f'_{p'}$ depends on the average of feature ϕ for both time series:

$$d_{rel}^{\phi}(f_p^\alpha, f_{p'}^\beta) = \left| \frac{\phi_{f_p^\alpha} - \phi_{f_{p'}^\beta}}{(\phi_{f_p^\alpha} + \phi_{f_{p'}^\beta})/2} \right|$$

□

Each feature is used either with the relative or the absolute distance (see Table 2.1). The category of each feature depends on the meaning. For instance, the deviation of the maximum consumption feature has to be considered in relation to the actual value. For households with a high maximum consumption value, small absolute differences may not be significant.

To re-identify a consumption record, we compare the feature value calculated on the known household $\phi_{f_p^\alpha}$ with the feature value of the record in question $\phi_{f_{p'}^\beta}$ by means of the absolute or relative distance. Because inaccuracies may occur even for the same household, we say that two feature values are equal if their distance is below a certain threshold δ_ϕ. To determine this threshold, we conduct the following steps: For each feature ϕ and each individual in $\mathcal{I} = \{p1, \ldots, pn\}$ we calculate a set of distances $D_\phi = \left\{d^\phi(f_{p1}^\alpha, f_{p1}^\beta), \ldots, d^\phi(f_{pn}^\alpha, f_{pn}^\beta)\right\}$. This set may contain outliers, e.g., an individual may go on vacation, leading to a huge difference in the consumed energy. To diminish the influence of outliers on the threshold, we take only the 'smallest' 90% of the values in D_ϕ, we call $D_\phi^{0.9}$ subsequently. In particular, $\left\|D_\phi^{0.9}\right\| = 0.9 \cdot \left\|D_\phi\right\|$ holds. We choose the following implementation of δ_ϕ.

Definition 17 (Implementation of δ_ϕ): Let $avg(X)$ be the average value of set X, and $SD(X)$ the standard deviation of the sample X. For each feature ϕ, the implementation of the tolerated error δ_ϕ is calculated as follows: $\delta_\phi = avg(D_\phi^{0.9}) + SD(D_\phi^{0.9})$. □

Finally, we are able to define the score assigned to two time series f^α and f^β depending on feature ϕ.

Definition 18 (Score for feature ϕ): The score $Score_\phi(f^\alpha, f^\beta)$ is the similarity between the time series f^α and f^β with respect to feature ϕ and is calculated as follows:

$$Score_\phi(f^\alpha, f^\beta) = \begin{cases} 0 & \text{if } d^\phi(f^\alpha, f^\beta) < \delta_\phi \\ \frac{d^\phi(f^\alpha, f^\beta) - \delta_\phi}{SD(D_\phi^{0.9})} & \text{otherwise} \end{cases}$$

□

The more similar two time series are with respect to a certain feature, the smaller is the score.

Weights Computation

An adversary is able to take a set of features Φ into account. A feature may be of higher importance than another for the specific data set. To support that case weights Ω are also considered for each feature. The resulting total score is calculated as follows.

Definition 19 (Score for set of features Φ and weights Ω): The score of two consumption records for a set of features is the sum of all normalized feature scores. We calculate the score as follows:

$$Score_\Phi(f^\alpha, f^\beta) = \sum_{\forall \phi \in \Phi} \omega_\phi \cdot Score_\phi(f^\alpha, f^\beta)$$

□

Our goal is to re-identify households by their consumption records in comparison to a training consumption record with known identity. Intuitively, to increase the re-identification performance, the weight for a distinct feature should be high and low for a less-distinctive feature. In total, we explain three different ways of determining these feature weights. They differ in their computing complexity. Thus, we investigate whether investing more computing time helps in improving the re-identification performance.

Static Approach: This is our baseline approach; all features are weighted equally with 1:

$$\forall \phi \in \Phi : \omega_\phi = 1$$

LP Approach: This approach uses linear optimization to determine weights that fulfill the following properties: Distances to other individuals should be maximized, whereas distances to the individual's own record should be minimized. We use linear optimization to maximize a term that iterates over each individual $p \in \mathcal{I}$ and totals the differences of the correctly re-identified record and the next closest individual p':

$$\sum_{\forall p \in \mathcal{I}} \left| min_{p' \in \mathcal{I} \wedge p' \neq p}(Score(f_p^\alpha, f_{p'}^\beta)) - Score(f(p)^\alpha, f_p^\beta) \right|$$

Obviously, this solution is not optimal with respect to the training data set: Maximizing the difference considering all individuals may still lead to a suboptimal solution for a single individual. In particular, maximizing that difference does not guarantee that the score to the individual's own record is the smallest.

ILP Approach: The ILP approach follows a similar intuition as the LP approach. The resulting score considering the data of the same individual has to be smaller than the score to any other. In contrast to the LP approach, we guarantee that the score to the individual's own record is the smallest if a valid solution is found. The binary variable $x_p \in \{0, 1\}$ indicates the following for $p, p' \in \mathcal{I}, p \neq p'$:

$$x_p = \begin{cases} 1 & \text{if } \exists p' : Score(f_p^\alpha, f_{p'}^\beta) \leq Score(f_p^\alpha, f_p^\beta) \\ 0 & \text{otherwise} \end{cases}$$

x_p is 0 if the score for the correct identification is smaller than for an incorrect one, and 1 otherwise. Thus, we have to minimize the following sum to get an optimal solution:

$$\sum_{p \in \mathcal{I}} x_p$$

To do so, we use linear optimization to determine the fitting weights.

Household Size	Fraction
1	40%
2	25%
3	20%
4	10%
5	5%

Table 2.2: Distribution of household sizes

2.2.3 Used Data Sets

To provide meaningful results, we base the evaluation for this chapter and throughout this thesis on real-world data. This section summarizes and explains contents of two data sources, used for this and for subsequent evaluations.

Electricity Customer Behavior Trial

The Irish Social Science Data Archive (ISSDA) publishes the 'Electricity Customer Behavior Trial' data set from the Commission for Energy Regulation (CER) in Ireland. The CER conducted a study with $5,000$ Irish homes of different sizes between 2009 and 2010 [51]. The households were equipped with a smart meter that measured the power consumption every 30 min. The data also included questionnaire results; however, in this work only the consumption data and the number of people living in a household are considered.

Distribution of Household Sizes

For the efficient computation of results, we usually consider only a fraction of the households in a huge data set. In particular, we define a town or a district by the total number of people living there. To (randomly) extract a representative set of household sizes, we require a distribution. National statistics offices publish such data, i.e., we took the data from the German Office to represent central Europe. Table 2.2 summarizes the statistics published [106].

2.2.4 Study Results

The goal of the study is to show that households can be re-identified with the help of rather simple features. Thus, the results are divided in two parts: First, we investigate the identifying degree of the features used; and second, the actual re-identification performance. Both results are computed on the CER data set. We vary the number of total individuals living in households of different sizes following a typical distribution (Both data sources are described in the previous Section 2.2.3).

Identifying Degree of Features

We consider a feature as identifying if the feature value itself is unique for the known period and if the difference to the same feature for the same individual in the re-identification period is low (optimally zero). More precisely, regarding the standard deviation of the feature values and differences, the following holds: A feature is identifying if the feature values have a ‘high’ standard deviation and the differences have relative ‘low’ value. As a first step, we investigate the features proposed on their identifying degree. For the evaluation, we take a set of households with 500 persons living in them. Figures 2.2–2.12 show the results represented as histograms. Table 2.3 lists the standard deviation of the feature values and the feature value differences.

R.1 *Features differ in the identifying degree.* Features differ in the standard deviation of the feature values and its differences. For instance, the feature ‘standard deviation’ has a lower deviation in the values than in the differences (Table 2.3). In turn, the feature ‘overall consumption’ has a 2.5-times higher deviation of the values compared with the differences. Thus, ‘overall consumption’ is more identifying than ‘standard Deviation’. The histograms (Figs. 2.2 and 2.7) confirm the results in the table.

The actual identifying degree of a single feature may also change when considering a different data set. However, the features are defined independently of an actual data set and thus are general.

R.2 *The features proposed are suitable for re-identification, i.e., they are sufficiently identifying.* Judging from the standard deviations in Table 2.3, there are a number of features with the deviation of values being a multiple of the differences. Large fractions of households having low differences and being well distributed over the range complement this fact. For example, this is the case for the feature ‘Consumption M–F, 9 p.m.–2 a.m.’ (Fig. 2.5). Because we consider a set of features for re-identification, the chosen features are sufficiently identifying.

We discuss the actual identifying degree in the next section, including further experimental analysis.

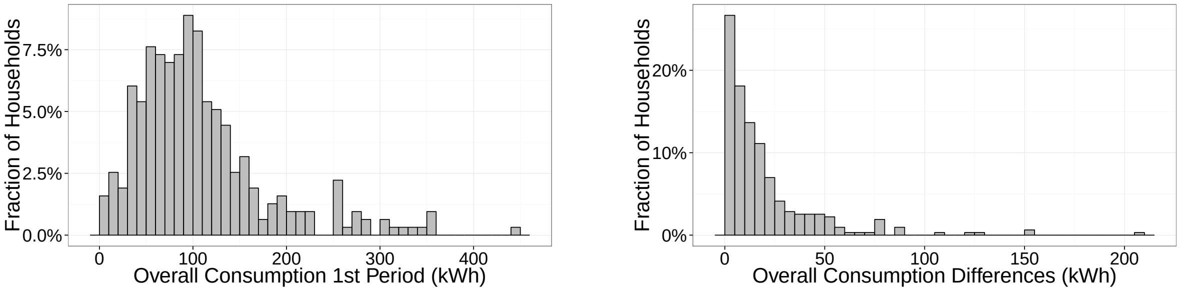
Re-identification Performance

To test the actual re-identification performance, we conducted experiments with the following varying parameters:

- We vary the total number of people distributed over the households among 100, 500, and $1,000$. This results in 32, 158, and 314 households, respectively. Half of the households are for training purposes ($\mathcal{F}^\alpha$), and the other half for testing the re-identification performance ($\mathcal{F}^\beta$).
- The considered timespan is 14, 21, and 28 days. The feature values of the first half of these timespans is used as external knowledge.
- For each configuration, we used the static as well as the LP approach for the weights computation. Calculating the weights with the ILP approach was, as a result of the computational

Standard Deviation of	Feature Values	Differences
Overall consumption	71.90	25.55
Maximum consumption	2.45	1.93
Minimum consumption	0.15	0.032
Consumption M–F, 9 p.m.–2 a.m.	14.98	0.41
Consumption M–F, 4 a.m.–8 a.m.	9.50	5.04
0.9-Quantile	1.09	0.47
Standard Deviation	0.43	1.44
Frequency of Mode	15.41	0.54
Weekend Consumption	0.06	0.058
Average Wakeup Hour	2.32	1.72
Average Bedtime Hour	5.72	3.39

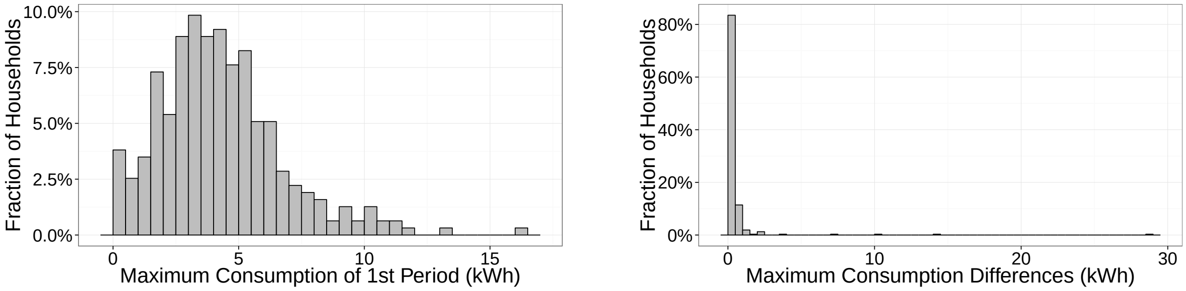
Table 2.3: Standard deviation of feature values and differences



(a) Distribution

(b) Differences Distribution

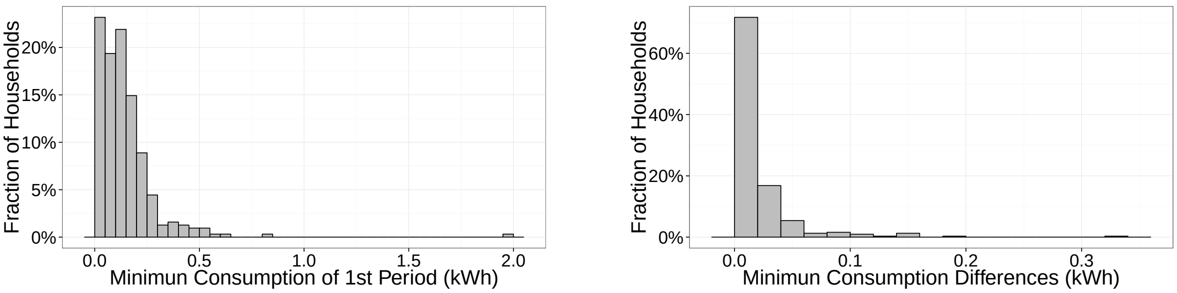
Figure 2.2: Analysis of the ‘Overall Consumption’ feature



(a) Distribution

(b) Differences Distribution

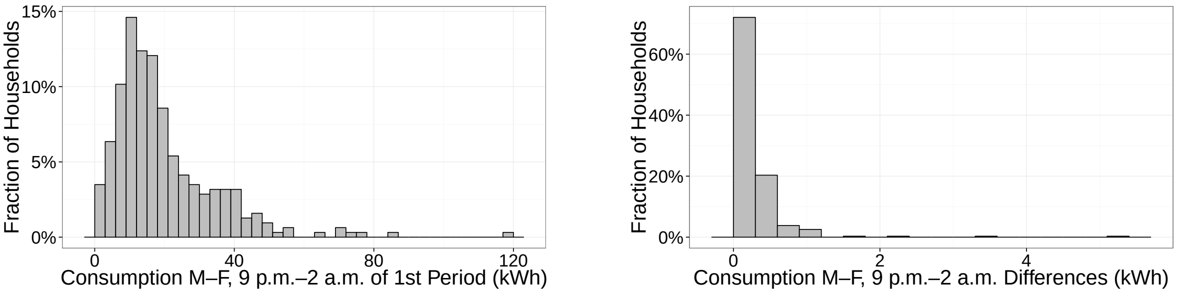
Figure 2.3: Analysis of the ‘Maximum Consumption’ feature



(a) Distribution

(b) Differences Distribution

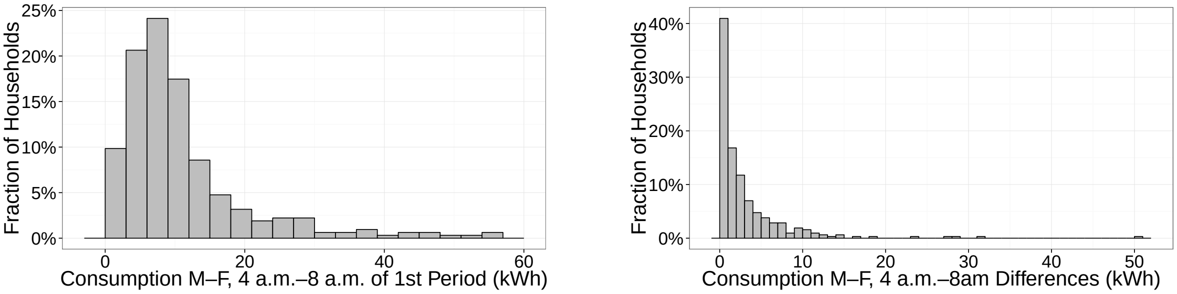
Figure 2.4: Analysis of the ‘Minimum Consumption’ feature



(a) Distribution

(b) Differences Distribution

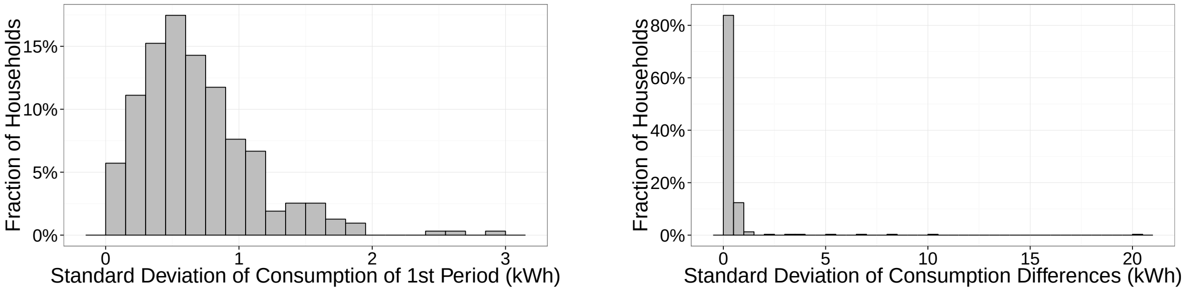
Figure 2.5: Analysis of the ‘Consumption M–F, 9 p.m.–2 a.m.’ feature



(a) Distribution

(b) Differences Distribution

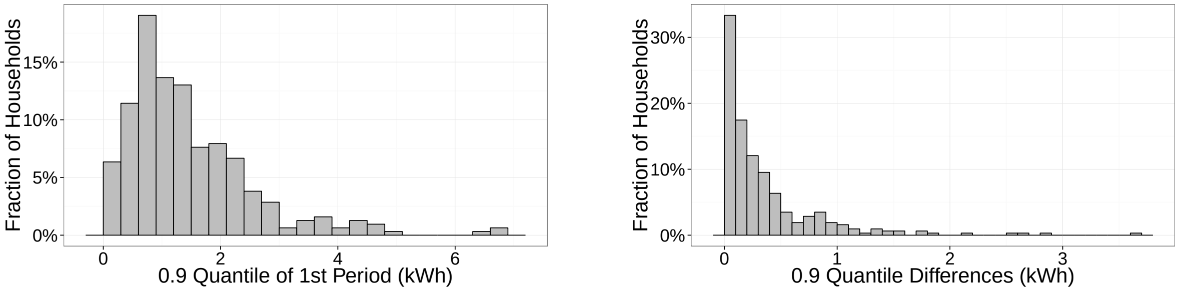
Figure 2.6: Analysis of the ‘Consumption M–F, 4 a.m.–8 a.m.’ feature



(a) Distribution

(b) Differences Distribution

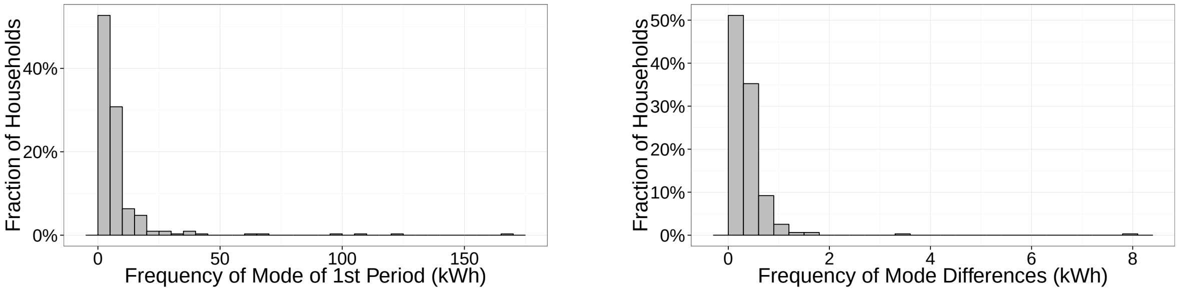
Figure 2.7: Analysis of the ‘Standard Deviation’ feature



(a) Distribution

(b) Differences Distribution

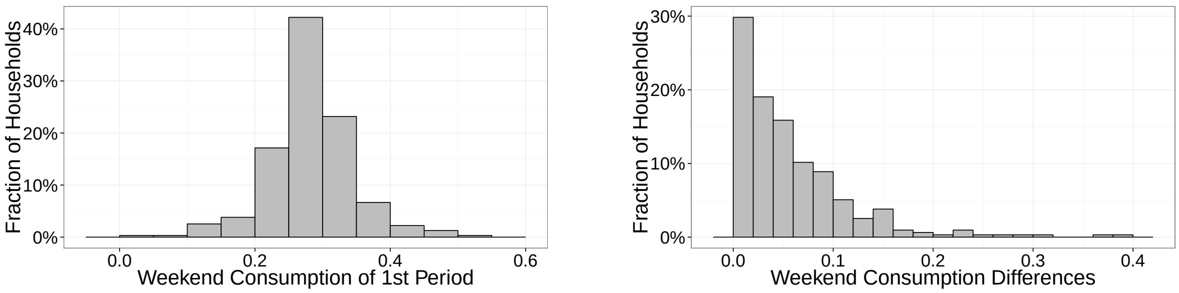
Figure 2.8: Analysis of the ‘0.9 Quantile’ feature



(a) Distribution

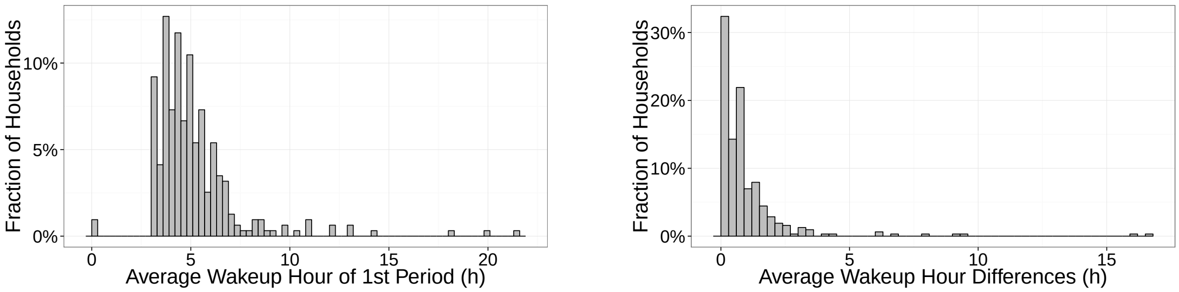
(b) Differences Distribution

Figure 2.9: Analysis of the ‘Frequency of Mode’ feature



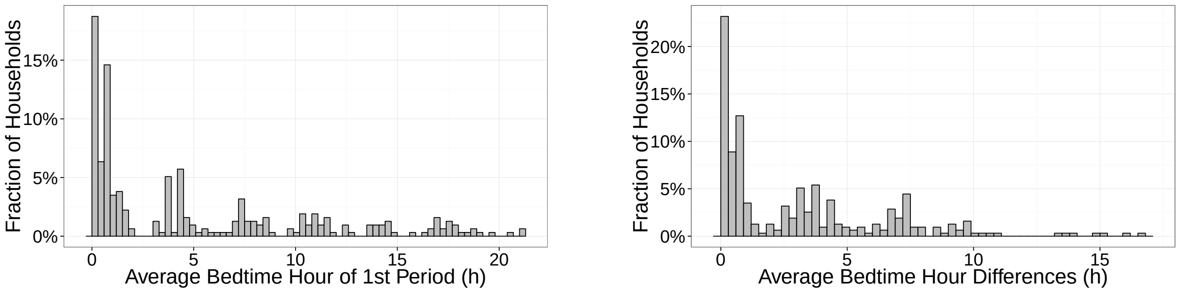
(a) Distribution (b) Differences Distribution

Figure 2.10: Analysis of the ‘Weekend Consumption’ feature



(a) Distribution (b) Differences Distribution

Figure 2.11: Analysis of the ‘Average Wakeup Hour’ feature



(a) Distribution (b) Differences Distribution

Figure 2.12: Analysis of the ‘Average Bedtime Hour’ feature

effort, only possible with 100 persons (32 households, respectively).

- Because the households are randomly selected, we repeat the training and the re-identification process for each configuration 10 times.

We measure the relative number of re-identified households, training, and test times for the different weighting functions. In particular, we calculate the re-identification rate.

Definition 20 (Re-identification Rate): The re-identification rate is the fraction of households of a test data set that can be re-identified. For the re-identification, we assume external knowledge $\mathcal{K}$ (according to Definition 4). □

R.3 *Re-identification rate decreases with the number of test households.* In general, if households have similar habits, they also have a similar power consumption. This makes re-identification more difficult. The larger the set of tested households, the more likely it is to have similar power consumption values (Fig. 2.13).

R.4 *Re-identification rate increases the longer the tested timespan.* The longer the tested timespan, the longer the timespan weights are trained. That leads to a more precise external knowledge, e.g., this reduces the probability that a feature is calculated on an atypical set of power consumption values for a household. Thus, the re-identification rate increases the longer the tested timespan (Fig. 2.13).

R.5 *Re-identification rate does not change much when using weights computed with the LP approach.* In theory, the LP approach tries to determine weights that increase the re-identification performance (Section 2.2.2). However, experiments show that the use of static weights is as good as using LP computed weights. Computing weights on the training data set $\mathcal{F}^{\alpha}$ gears the weights to exactly this data set, but this does not necessarily increase the re-identification rate on the test data set $\mathcal{F}^{\beta}$.

R.6 *The ILP approach has a higher re-identification rate compared with the LP approach, but a similar rate compared with the static approach.* The ILP and the LP weights computation approaches strive to increase the re-identification performance. However, we cannot guarantee that because the training and test data set may have varying characteristics. In Result **R.5** we have already shown that the LP approach cannot compete against static weights. The ILP approach results in higher re-identification rates than the LP approach, but is head-to-head with the static approach (Fig. 2.14).

In addition the re-identification rate, we investigate whether investing additional computational effort in the training phase with the LP or ILP approach increases re-identification success. We measure execution times on a machine with a Dual-Core AMD Opteron 2218 Processor at 2,600 MHz and 28 GB RAM. The code is implemented in Java and runs on a Java VM 1.7. For the equation solving (LP and ILP approach) we use the lp_solve[1] 5.5 library.

[1] http://lpsolve.sourceforge.net/

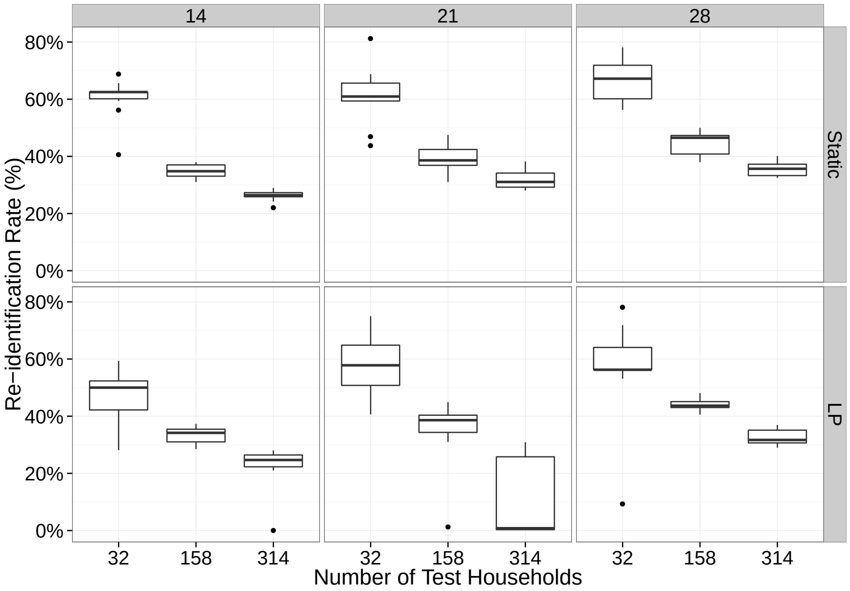


Figure 2.13: Re-identification rate for different tested timespans, different weight computation approaches, and different household sizes

R.7 *The training time of the the LP is at least an order of magnitude longer than for the static approach.* The training phase of the static approach includes only the computation of the tolerated errors δ_ϕ. The training phase of the LP approach additionally includes the computation of weights following the required properties. Solving these equations naturally consumes time. The measured times (Fig. 2.15) show that the training time of the LP approach takes at least 10 times longer. The differences in the execution times also increase with the number of training households.

R.8 *The ILP approach requires longer computation time than the LP approach.* Because of the computational complexity, we were only able to conduct experiments with the ILP approach covering the 100 persons (32 individual households, respectively) scenario (Fig. 2.16). The average training time of the ILP approach is a few seconds longer than the one of the LP approach. However, improvements compared with the performance of the static approach are rather small. Thus, the additionally invested time does not lead to a significant increase in performance.

The increased complexity is acceptable if in turn the re-identification rate increases. However, at least for the investigated data sets, this is not the case.

R.9 *The additional computational effort for the LP and the ILP approach does not lead to a similar significant increase in the re-identification rate.* Results **R.5** and **R.6** show that the LP approach may lead to worse re-identification rates, whereas the ILP approach may increase the

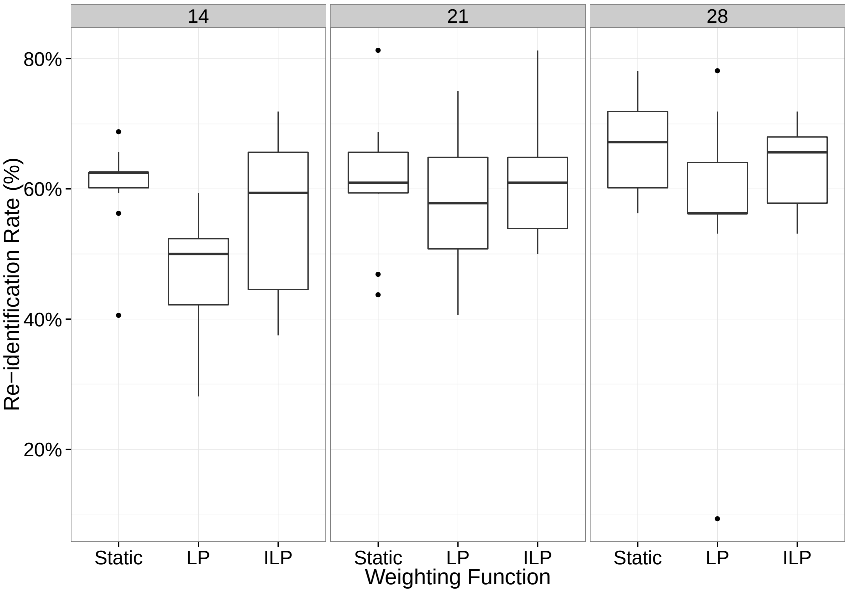


Figure 2.14: Re-identification rate for the different tested timespans and different weight computation approaches including the ILP approach for 32 households

re-identification performance in certain situations. However, the small increase in re-identification is not proportional to the increase in computation time.

Summarizing all the findings in the context of re-identification, we come to the final result.

R.10 *Smart-meter data can be re-identified.* With a minimal computational effort it is possible to re-identify up to almost 80% of the households in the test data set (Fig. 2.13). For certain data sets it is even possible to increase the re-identification rate by investing more in the training time, i.e., using the ILP approach (Fig. 2.14). Shorter tested timespans and a larger number of households decreases the performance; however, the re-identification rate never drops below 30% on average. Keeping this in mind, without the proposed re-identification method, one may simply guess the right household. This would lead to re-identification rates ranging between 0.3% for 1,000 individuals up to 3% for the scenario with only 100 individuals. The achieved rates are by an order of magnitude higher, so we can conclude that households can be re-identified.

2.3 Conclusions

In this chapter, we have seen that the time series of smart-meter data contain different personal information that reflects the living conditions. Depending on the actual frequency of measurements, different details about the household habits can be extracted from such time series up to the running

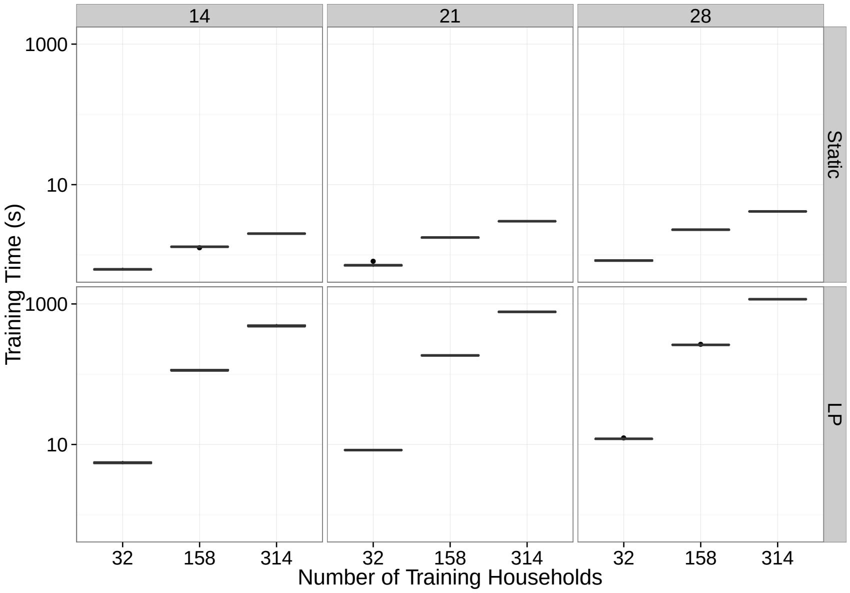


Figure 2.15: Training time for different tested timespans and different weight computation approaches

TV program. The conducted study clearly shows that it is possible to re-identify households while having certain external knowledge. Thus, the data can be referenced to individuals and are subject to privacy legislation. In particular, the whole approach—including feature extraction, training, and re-identification phase—requires minimal computational effort and minimal personal effort as well. Re-identification of smart-meter data threatens the privacy of a rising share of the population. As explained in Section 1.2.1, legislation and national interests foster the deployment of smart meters. Additionally, smart grid applications process such data.

To protect the individual right of informational self-determination while still providing applications with personal data, we require a method for removing information deemed private before publication. The proposed features for re-identification and the extractable information using methods explained in Section 2.1.1 might be an indication of the kind of information that is private. In the following chapter, we explain our second contribution, a privacy-enhancing method for the time series of smart-meter data.

The proposed features and evaluation is geared toward smart-meter data. This is necessary to have reliable evaluation results. However, with different feature definitions, the proposed method is applicable to other kinds of time series as well. From an intuitive perspective, it is very likely that there are periodic events or repeating patterns in personal time series, regardless of the actual data source, because daily life regularly follows periodical routines. Thus, the conclusions drawn

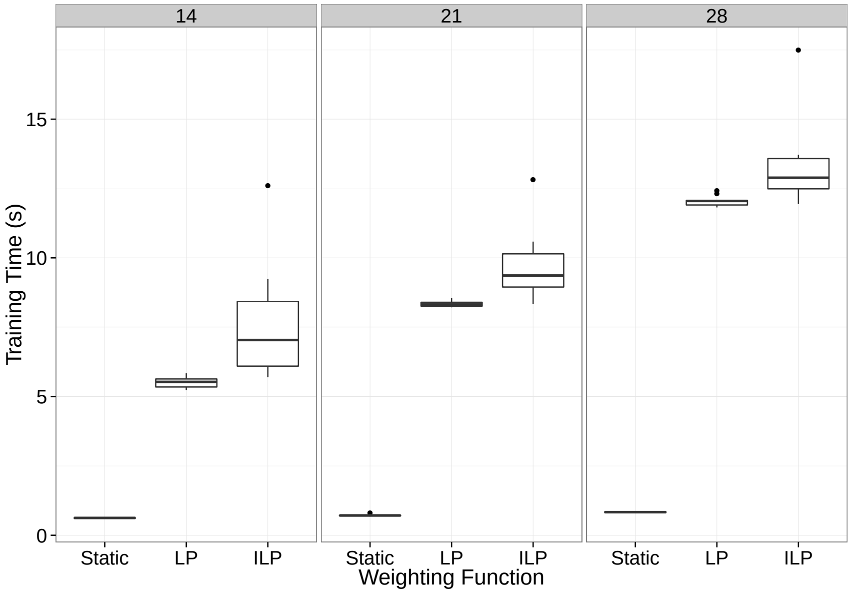


Figure 2.16: Training time for different tested timespans and different weight computation approaches including the ILP approach for the 100 persons/32 households scenario

in this chapter are also applicable to time series in general.

Chapter 3

Privacy-Enhancing Methods

In many domains the disclosure of personal data is important to facilitate innovations. Hence, it is of common interest to provide access to these kinds of data to achieve social goals. In contrast, individuals require privacy protection. This antagonism holds for the smart grid (see Chapter 1). In particular, the smart grid strives to reduce CO_2 emissions and to guarantee reliable supply at affordable prices. To achieve both goals it is necessary to integrate renewable energy sources into the electrical grid. In contrast to traditional energy power plants, most of the renewable sources cannot provide a steady and controllable supply. The installed smart meters in private households play a major role in the integration. They allow a two-way communication between the producers and consumers, thus improving the coordination of supply and demand. Applications like demand-side management, flexible tariffs, or the local energy market improve the efficiency of the power distribution. All of these examples have in common that they require access to the power consumption data of individual households.

Smart-meter data contain a variety of different personal information. ‘Personal’ in a privacy context means that it reflects the living conditions of individuals (see Section 1.2.2). In general, personal information is subject to privacy legislation and requires protection as long as the data can be referenced to a single individual. However, power-consumption data is easily identified, as we have seen in Chapter 2. Additionally, applications like demand-side management require data with identifiers. A complementary approach is to remove the information that an individual deems sensitive, which we refer to as ‘secrets’ subsequently. If such information is removed, individuals might be willing to publish their data because their right of informational self-determination is respected.

In this chapter, we focus on the time series of smart-meter data. However, most of the ideas are also applicable to time series in general. For a discussion, see Section 3.5. The contents of this chapter are published in [58].

Example 2 ***(Bob’s electric flow heater)*****:** Bob has a smart meter and heats his showering water with an electric flow heater. Bob will accept the disclosure of his smart-meter data if certain private information is removed. In particular, he wants to keep private the time of when he is showering on the weekends and during the weekday mornings. An adversary with access to the published data should not be able to learn whether the flow heater is starting or stopping between 8:00 and 11:00 on a weekday by inspecting the (disclosed) smart-meter time series. On weekends, the consumption data should be so noisy that the probability of inferring the exact time when the heater is working is sufficiently low. To do so, one has to know how the time series reflects the flow-heater use and hide this on a weekday and then detect when the flow heater starts and stops on a weekend. The noise should not be excessive, i.e., to preserve utility, the data should still contain information that Bob does not explicitly want to hide. □

Smart-meter data contain a lot of different personal information, as stated in Chapter 2. Thus, secrets may differ for each individual and respecting the privacy preferences means giving certain guarantees to which extent the information is actually removed or hidden. The well-known Pufferfish privacy framework [60] supports the definition of individual and understandable privacy preferences and their semantics. It also covers correlations within the data set, which is sometimes necessary to guarantee privacy while keeping utility. Differential privacy in turn leaves aside such correlations.

Example 3 ***(Correlations in the data)*****:** Let $f(A), f(B), f(C)$ be the smart-meter time series of Alice, Bob, and Carl's household. $f(A)[t]$ is the total power consumption of Bob's household at time slot t. Differential privacy approaches [5, 99] publish the privacy-enhanced sum at each time slot of the households considered, i.e., $f(B)[t] + f(A)[t] + f(C)[t] + \ldots$: If there are no correlations about the consumptions of Bob, Alice and Carl, an adversary cannot infer the actual consumption of any one of them. However, there also are correlations when looking at each time series in isolation: Suppose that Alice, Bob, and Carl each have a flow heater (for the shower) and bath lighting. $f(B)^1[t]$ is Bob's flow heater consumption and $f(B)^2[t]$ the one of the bath lighting. $f(B)[t]$ is the sum of all appliances in Bob's household: $f(B)[t] = f(B)^1[t] + f(B)^2[t] + \ldots$. Privacy cannot be guaranteed in the same way as for the sum of $f(B)[t], f(A)[t]$ and $f(C)[t]$: The flow heater and the bath lighting obviously have correlations that differential privacy does not address [59]. □

Certain sensitive information is guaranteed to be removed from the data set. Pufferfish is an abstract framework that has not yet been applied to smart-meter data. Thus, a qualitative evaluation of the resulting data set is also missing. The application requires challenging conceptual work: We require an abstracted representation of private information in a time series of smart-meter data to perturb the data set according to abstract Pufferfish guarantees. All of the approaches have to be general enough to cover arbitrary privacy preferences. We need to measure the utility of the resulting privacy-enhanced data set. In the following, we elaborate more on these challenges.

Representation of Private Information Because of the nature of smart-meter data, the aggregated power consumption of the running devices is reflected. For the sake of simplicity, 'devices' in this context summarizes all electric power-consuming appliances, including lighting. We assume that certain activities are related to a set of running devices. Running devices, such as the flow heater in Example 2, result in a sequence of power-consumption values. These values may vary because devices may have a slightly different power-consumption characteristic each time they run. The exact consumption development may depend on external factors like the temperature. Additionally, the smart meter may measure the aggregated power consumption of several running devices. The first challenge is to find an abstracted representation that is flexible enough to cover the explained uncertainty, and precise enough to have meaning for the secrets in question. In the following, we refer to a single value of such an abstracted representation as 'coefficient'. Furthermore, we require the abstraction to have clear-cut semantics and the transformation must be well-defined. We explain the requirements in more detail in Section 3.2.1. The overall goal of the transformation is to have a representation of the time series in which each coefficient has a meaning pertaining to a certain secret. This is in contrast to the natural time-based representation, in which each coefficient has a meaning for a point of time.

Example 4 ***(Bob's flow heater, abstraction, and coefficients)*:** In Example 2, Bob wants to hide the activity of his flow heater. Thus, the coefficients have to allow conclusions regarding the flow heater operation. Suppose that the flow heater consumes 25 kW when running, and 0 kW otherwise. A starting flow heater will lead to a difference between the power consumption at point of time t and at $t+1$ of 25 kW. Such a difference indicates when Bob starts showering, and this is subject to his privacy requirement. An abstracted representation in which each coefficient reflects this kind of change is appropriate for the flow heater example. In general, running appliances lead to more complex developments. For instance, a washing machine has different cycles with changing electricity demand. If this is relevant to someone's privacy, this information must be abstracted and then hidden. □

Perturbing Smart-Meter Data Pufferfish provides precise guarantees for user-defined secrets. Usually privacy is achieved through perturbation. Applying noise to the time series of smart-meter data, however, is not straightforward: Such data are usually an aggregate of several appliances and require a decomposition on a conceptual level. Next, we must take into account that different appliances in the decomposed representation may have correlations. Our objective is to deal with such time series individually.

Generality Secrets require a specific abstracted representation to achieve Pufferfish privacy. It is challenging to find abstracted representations for a wide range of privacy requirements.

Evaluation Quantifying the usefulness of data is also not obvious. First, the data quality rating requires a meaningful set of privacy requirements. In this chapter, we provide an objective on a realistic source of such requirements. Second, we require an application-specific measure to quantify the actual effect. We will discuss the measure in Chapter 4, and the results in Chapter 5.

Contributions We contribute PACTS (Contribution **C.2**), a provable privacy-enhancing method for time series. PACTS addresses these challenges as follows: The variety of possible secrets is broad. To cover them we carefully select different abstracted representations and their transformations. To illustrate the whole method, we use the wavelet transformation as an example; it already covers different secrets. To ensure privacy according to ϵ Pufferfish privacy, PACTS decomposes the abstracted aggregated smart-meter signal into several channels on a conceptual level. This decomposition allows the application of ϵ Pufferfish privacy-conform noise. Before publishing the time series, it is transformed back into the natural, time-based representation. Thus, the published privacy-enhanced and the original time series have the same format.

To ensure generality, we have to show that the transformation step is capable of covering a wide range of objective privacy requirements. To do so, we take recent information-extraction methods as an objective source of possible secrets. In particular, we define secrets covering re-identification [11] (Chapter 2) and a non-intrusive-appliance-load monitoring [8] approach for information extraction.

The structure for the remainder of this chapter is as follows: First, we discuss several fundamentals (Section 3.1), including related privacy-enhancing approaches and the requirements for the proposed PACTS approach (Section 3.2). We evaluate the effectiveness against information extraction as well as re-identification (Section 3.4) before we conclude (Section 3.5).

3.1 Fundamentals

First we define a common notation in Section 3.1.1 and then we review recent privacy approaches (Section 3.1.2). PACTS relies on the Pufferfish privacy framework, which we introduce in detail in Section 3.1.2. Throughout this chapter, we use the wavelet transformation as an example in PACTS. Thus, we introduce wavelets in Section 3.1.4. Please not that, PACTS is not limited to this single transformation.

3.1.1 Notation

One of the key elements of this approach is the transformation of a time series in an abstracted representation. For the sake of an intuitive illustration, we have chosen a vector-based representation in the context of this chapter. Vectors are elements of a vector space. The coefficients of each vector are defined on a basis and express a linear combination of the vectors contained in the basis. This basis defines the meaning of the vector coefficients. The standard representation of a time series is still related to certain points of time and a value domain, i.e., the measured power-consumption values. In general, time series are infinite sequences of measured values related to an also infinite number of points of time. The support of vectors with infinite length requires the definition of an infinite basis. However, handling infinite bases and vectors makes the illustration of the transformations and abstracted representations unnecessarily complex. Thus, we make the nonrestrictive assumption that the basis of the considered vector spaces are finite. The proposed notation of the time domain (Notation 1) and the time series (Notation 3) includes an infinite number of measurements. Thus, we redefine both.

Notation 5 (Finite Time Domain): $\widehat{\mathcal{T}}$ is the standard time domain. We assume that it is discrete and of finite length: $\|\widehat{\mathcal{T}}\| < \infty$.

Notation 6 (Time Series in Vector Representation): A time series in vector representation is an n-dimensional vector with basis B referred to as f_B. If we specifically refer to coefficient t we denote that as $f_B[t]$.

As previously mentioned, the standard basis is still the canonical time basis. To easily distinguish between the standard and the abstracted representation, we denote the standard basis as E, defined as follows.

Notation 7 (Time Series Standard Basis): The basis E is the standard basis for time series in vector representation. Consequently, if the basis of a time-series vector is not explicitly given, it is the standard basis: $f = f_E$

The standard basis E maps the time domain $\widehat{\mathcal{T}}$ to a time-series vector as follows. Let $[t_1, \dots, t_n]$ be the ordered list of all elements $t_i \in \widehat{\mathcal{T}}$, then $f_E[t_i] = f_E^{\intercal} \cdot e_i$ is the power consumption at time slot t_i. For a given $\widehat{\mathcal{T}}$, the basis vector e_i represents the ith ordered element. Consequently, the standard basis is $E = \{e_i | i = 1 \dots n\}$. Additionally, we assume that the measured time series is discrete and the intervals between the measured points are of equal length.

Notation 8 (Vector Space): $\mathcal{V}_B$ is the vector space containing all linear combinations of the basis elements in B.

The presented and related approaches define a privacy mechanism. In general, those mechanisms are defined as follows:

Definition 21 (Privacy Mechanism $\mathcal{M}_p$): A privacy mechanism $\mathcal{M}_p(\mathcal{D}ata)$ takes a data set $\mathcal{D}ata$ and a parameter set p and returns a privacy-enhanced representation of $\mathcal{D}ata$. □

3.1.2 Privacy Protection Approaches

Privacy-preserving data publishing has been intensively studied in literature. To give an overview, we group similar approaches in different categories. In terms of privacy enhancement, we distinguish between anonymization and perturbation. Anonymization approaches strive to hinder an adversary from linking data to individuals. Without this connection, the data are not longer personal (see Section 1.2.2). Perturbation, in turn, hides specific privacy-relevant or sensitive parts of the data, usually by adding systematic noise. Records in the resulting privacy-enhanced data set may still refer to single individuals, but ideally do not contain the sensitive information anymore. We assume that there is a trustworthy third party, if necessary, for the computation of the privacy-enhanced result. Instead of focusing on the privacy-aware aggregation of several measurements, like in [40], we assume that privacy enhancing can be computed without any threat.

For the data quality measurements (Chapter 4) and the evaluation (Chapter 5) we choose one

Name	ZIP Code	Date of Birth	Disease
Paul	76131	7.30.1975	bronchitis
Martin	76351	10.13.1978	angina
Albert	76131	1.17.1970	leg fracture
Alice	68159	9.8.1987	breast cancer
Vanessa	68159	10.13.1980	cough
Bob	10115	5.2.1964	cold

Table 3.1: Non-anonymized patient records

representative, related method for each class. For the anonymization approach we choose [86], and for the perturbation approach we choose [88]. We explain both methods in greater detail in the following.

k-Anonymity and Extensions Based on the finding that removing direct identificators is not sufficient to guarantee anonymity [22], Sweeney et al. defined the notion of k-anonymity [108]. In a data set, we distinguish between identifiers, like names, quasi-identifiers which may lead to identification if an adversary has external knowledge and the sensitive part of the actual data. A data set is k-anonymous if at least k records have the same quasi-identifier. An adversary having external knowledge on the quasi-identifier cannot distinguish the actual individual being searched for among k. k-Anonymity usually is achieved by suppressing or generalizing quasi-identifiers. For example, the patient records in Table 3.1 consist of the 'Name' as an identifying attribute, the 'ZIP Code' and the 'Date of Birth' as quasi-identifiers, and the 'Disease' as a sensitive attribute. First, we remove the 'Name' column to strip the identifiers. However, an adversary having external knowledge regarding the residence or the age of an individual can easily re-identify a single record. In turn, Table 3.2 lists the $k = 2$ anonymous representation of the patient records. An adversary with the same external knowledge cannot single out a record because at least two records are indistinguishable with respect to the quasi-identifiers. However, attacks are still possible. First, the process of building groups of k records does not consider the content of the sensitive attribute. In our example this may lead to k records having the same 'Disease'. An adversary that is able to link an individual to such a group can easily reveal the disease. l-Diversity [74] takes care of this attack by requiring different values of the sensitive attribute in a k-group. Another attack is possible if the distribution of the sensitive attribute within the k-group differs much compared with common knowledge, e.g., if all records in a k-group have a different but very seldom lung disease. This allows an adversary to gain knowledge even though the actual disease cannot be extracted, i.e., that the individual in question has a seldom lung disease. To overcome this, t-closeness [69] matches the distributions of common knowledge with the distributions inside a k-group.

ZIP Code	Date of Birth	Disease
76***	197*	bronchitis
76***	197*	angina
*****	19**	leg fracture
68***	198*	breast cancer
68***	198*	cough
*****	19**	cold

Table 3.2: $k = 2$ anonymous representation of 3.1

k-Anonymity Derivatives on Time Series The k-anonymity principle also got attention in the context of time series. In a relational data set such as the patient records example (Tables 3.1 and 3.2), the distinction between quasi-identifiers and sensitive data is clear. However, in the context of time series, a common case is that the quasi-identifiers and the sensitive attribute are the same. Remember the re-identification approach in Chapter 2: Features of the time series itself lead to re-identification. Thus, building groups of k individual records requires the modification of sensitive data. In particular, each time series has to be indistinguishable compared with $k - 1$ others. Being indistinguishable means that points of time and the actual values are the same. For example, the canonical solution would be to compute the average values of all elements in a k group. However, the data can also be generalized to a certain range or outliers could be suppressed.

Similar to hiding trajectories of moving objects, a number of approaches exists that implement k-anonymity on such time series. The considered time and location dimensions allow different optimization of creating a k-anonymous data set. [1] tries to preserve information on the location dimension, whereas the successor [2] focuses on the time dimension. Instead of generalization, the data can also be suppressed [20] or randomized [86] to achieve k-anonymity. Depending on the assumptions regarding the external knowledge of an adversary, the quasi-identifier definition is not straightforward and may result in possible privacy breach. A different model, that is geared toward moving object data is presented in [120]. From a data-mining perspective, it is often necessary to extract the patterns present in such time series. Approaches like in [105] explicitly strive to protect these patterns. All of these approaches create a publishable data set. [44, 82] propose a trustworthy third party that ensures k-anonymity for specific user queries, e.g., [82] makes sure that the use of location-based services is anonymized to the service provider. An approach geared toward smart-meter data [56] tries to preserve as much information as possible by allowing the user to define an upper bound for information exposure. The upper bound is a number of points of time that an adversary is allowed to infer, and at all the other times a time series is indistinguishable among $k - 1$ others.

The common goals of all k-anonymity-based methods are to modify the values of time series so that the sequence of time/value pairs from one individual is identical to sequences of $k - 1$ others, and the anonymized time series is as similar to the original one as possible. An example of a similarity measure is the L2 norm:

$$dist(f_p, f'_p) = \sqrt{\sum_{t \in \mathcal{T}} \Big(f_p(t) - f'_p\Big)^2}$$

Because creating a k-anonymous database with minimal changes is NP-hard [6], we present Algorithm 1, which relies on a well-known heuristic [86]: First, it randomly selects a time series f_p from one individual. Next it chooses $k-1$ time series from other individual that has the smallest distance to the first time series, e.g., in terms of the L2 norm. In particular, method $selectGroup(k, p, \mathcal{I})$ computes a set of individuals $\mathcal{P}$ so that $\mathcal{P} \subseteq \mathcal{I}$, $p \in \mathcal{P}$, $|\mathcal{P}| = k$, and $max_{i \in \mathcal{P}}(dist(f_p, f_i)) \leq min_{j \in (\mathcal{I} \setminus \mathcal{P})}(dist(f_p, f_j))$. The algorithm repeats until fewer than k time series are not assigned to a k-group. They are assigned to existing k-groups by method $findGroup(\mathcal{K}, p)$, which returns a k-group $\mathcal{J}$ from a set of k-groups $\mathcal{K}$ so that $\mathcal{J} \in \mathcal{K}$ and $max_{j \in \mathcal{J}}(dist(f_p, f_j)) \leq min_{i \in (\bigcup(\mathcal{K} \setminus \mathcal{J}))}(dist(f_p, f_i))$. Finally, the algorithm replaces $f_p(t)$ with the group average for each point of time.

Algorithm 1: $\mathcal{M}_k$ Implementation for k-Anonymity

```
Data: $k$,Time domain $\mathcal{T}$,Set of individuals $\mathcal{I}$, Set of time series $\mathcal{F}$
Result: Anonymized set of time series $\mathcal{F}'$
Set kGroups = $\varnothing$;
// create groups of $k$ similar time series
while $|\mathcal{I}| \geq k$ do
    Individual $p$= $\mathcal{I}$.getRandomly();
    Set s = selectGroup($k$, $p$, $\mathcal{I}$);
    $\mathcal{I}$.subtract(s);
    kGroups.add(s);
end
// assign the remaining time series
foreach $p \in \mathcal{I}$ do findGroup(kGroups, $p$).add($p$) ;
// anonymize value for each k-group
foreach $kGroup \in kGroups$ do
    foreach $p \in kGroup,\ t \in \mathcal{T}$ do
        $f'_p(t) = \frac{\sum_{j \in kGroup} f_j(t)}{kGroup.size()}$;
        $\mathcal{F}'$.add($f'_p(t)$);
    end
end
```

The described k-anonymous derivatives have the following major drawbacks. First, considering a specific time series, the anonymized result is influenced by the user-defined parameters and the

Algorithm 2: $\mathcal{M}_\sigma$ Privacy-Enhancement Method Using Wavelet Transformation [88]

```
Input: Privacy Parameter $\sigma$
Input: Set of time series $\mathcal{F}$
Result: Privacy-Enhanced time series $\mathcal{F}'$
foreach $f_p(t) \in \mathcal{F}$ do
    $\widetilde{f_p(l,t)} = DWT(f_p(t))$ //Wavelet transform;
    $I_l = \left\{t : \left|\widetilde{f_p(l,t)}\right| \geq \sigma\right\}$;
    foreach level $l$ do
        $K = \sum_l K_l$ //coeffs exceeding $l$;
        $p = |N|/K$ //Noise ‘density’, N is number of coefficients;
        foreach detail $\widetilde{f_p(l,t)}$ do
            if $t \in I_l$ then
                $\widetilde{f_p(l,t)} + = GaussRnd(0, \sigma\sqrt{p})$;
            end
        end
    end
    $f'_t(=)InvDWT(\widetilde{f_p(l,t)})$;
    $\mathcal{F}' = \mathcal{F}' \cup \{f'_t()\}$;
end
return $\mathcal{F}'$;
```

remaining data set. For example, if all time series are similar, the resulting generalized time series will not differ much from the originals. Hence, we cannot guarantee that a certain piece of information described by the time-series development is removed or not. Second, the parameters, i.e., the k, require a global definition for the whole data set. Individually defined parameters are not applicable. Thus, in a scenario that requires custom-definable privacy requirements with guarantees, those methods are not suitable.

Perturbation Approaches To support individual privacy preferences, one possibility is to add systematic noise to time series in isolation. Uncorrelated noise applied to a time series is easily filtered out by means of wavelet-based filtering [27, 28]. To circumvent this, we need to apply noise that is dependent on the wavelet or fourier representation of the actual time series [88]: Let K be the number of wavelet coefficients exceeding σ, and N the total number of wavelet coefficients. Next, noise with the standard deviation of $\sigma \cdot \sqrt{N/K}$ and the mean value is the current coefficient if it is greater than or equal to σ. See Algorithm 2 for a pseudo code implementation of the wavelet-based perturbation. According to [88], this method ensures that only a small percentage of noise can be removed.

However, the data owners cannot decide *what* exactly is perturbed. Information may unnecessarily be perturbed and sensitive information may still be present.

Smart Meter–Specific Approaches Protecting the privacy of individuals in smart-meter data transferred to the utility company is possible with the help of a rechargeable battery: Consuming power directly from the house connection results in smart-meter data that contain possible sensitive information. Taking electricity instead from a locally installed rechargeable battery, leads to smart-meter data that reflect load cycles only [54, 96, 112]. An additional power router decides which source of electricity is chosen, guarantees security of supply, and is responsible for the battery management. However, privacy is bound to the battery capacity, e.g., if a specific activity requires more power than the battery currently can provide, the system is unable to hide this.

Provable Privacy Approaches Differential privacy [29] is an abstract and strict notion of guaranteed privacy for statistical databases, defined as follows: Adding or removing the data from a single individual from the data set does not significantly change the output. In other words, the record of a single individual has limited influence on the output. In turn, an adversary cannot gain information about a single individual. It has been applied to smart-meter data [5] and time series [99]. Example 3 illustrates the limitations. Additionally, it is especially challenging to provide utility and provable privacy guarantees, because in many cases the resulting data are not useful for applications anymore [75]. [118] shows that using wavelet transformation in combination with differential privacy preserves utility of range queries on a relational data set. Instead of applying noise to each entry, [118] perturbs the wavelet-transformed frequency matrix, thus reducing noise for count queries while still preserving individual privacy. However, this method is specifically geared to range queries on a relational data set and thus is not applicable to time series in general.

Theoretical results show that providing both provable privacy guarantees and utility is only possible if assumptions regarding the external knowledge of an adversary and the original data are made [59]. The Pufferfish framework [60] supports such assumptions in an abstract way and lets individuals define their privacy preferences. It has not yet been implemented and evaluated on smart-meter data. PACTS is such an implementation. Thus, we explain Pufferfish in greater detail in Section 3.1.3.

3.1.3 ϵ-Pufferfish Framework

The ϵ-Pufferfish framework [60] is a generalization of differential privacy [29], providing provable privacy guarantees while preserving utility. It is based on the theoretical ideas in [59]. In contrast to differential privacy, Pufferfish allows us to make assumptions regarding the adversary. To do so, probability distributions called data evolution scenarios $\mathcal{D}$ describe external knowledge on how the data were created. Furthermore, the set of potential secrets $\mathcal{S}$ describes *which* information can be hidden, but does not necessarily have to. $\mathcal{S}$ is a domain for $\mathcal{S}_{pairs}$ that contains pairs of secrets describing *how* a piece of information should be hidden.

Examples of secrets contained in $\mathcal{S}$ for the relational model are 'Bob has cancer.' There is no limitation for the abstraction level of such secrets, e.g., 'The record of individual i is in the data set' is also a possible secret. Secrets are facts that an individual wants to hide. However, a single secret s does not define 'how' the specific information should be hidden. $\mathcal{S}_{pairs}$ is a subset of $\mathcal{S} \times \mathcal{S}$, which defines what an adversary should not be able to distinguish. A canonical example in $\mathcal{S}_{pairs}$ would be ('Alice has cancer.','¬Alice has cancer.'). Pufferfish allows arbitrary pairs (s_i, s_j) to specify precisely the information to be hidden. Continuing our example, Alice may only require hiding which kind of cancer she has. In that case the discriminative pair is ('Alice has lung cancer.','¬ Alice has stomach cancer.'). This is advantageous because hiding specific secrets like the kind of cancer tends to require less noise than hiding general secrets like having cancer at all. The only requirement for discriminative pairs is that they have to be mutually exclusive but not necessarily exhaustive, i.e., at most, one is true but both can be false.

Data-evolution scenarios $\mathcal{D}$ contain assumptions on how the data have been generated. This is external knowledge of an adversary. It is a set of probability distributions over possible database instances that quantify how likely a certain fact is. For instance, the probability that a certain patient has cancer is higher for a hospital that is a cancer center compared with the data set of a general hospital. Each distribution $d \in \mathcal{D}$ corresponds to the external knowledge of an adversary on how the data have been generated. For example, $P(\mathcal{D}ata = \{x_1, ..., x_n\} | d_f) = p(x_1) \cdot ... \cdot p(x_n)$ if the probabilities of each record in $\mathcal{I}$ are independent. $P(\mathcal{D}ata = \{x_1, ..., x_n\} | d_p)$ is the conditional probability that $\mathcal{D}ata$ is $\{x_1, ..., x_n\}$ under d_p.

Furthermore, a privacy mechanism $\mathcal{M}$ is a method for transferring a data set $\mathcal{D}ata$ into a privacy-enhanced representation $\mathcal{M}(\mathcal{D}ata)$ (see Definition 21). It guarantees ϵ-Pufferfish privacy if it fulfills the following definition [60]:

Definition 22 (ϵ-Pufferfish privacy): Given a set of Secrets $\mathcal{S}$, a set of discriminative pairs $\mathcal{S}_{pairs}$, data-evolution scenarios $\mathcal{D}$, and a privacy parameter $\epsilon > 0$, a privacy mechanism $\mathcal{M}_\epsilon$ satisfies ϵ-Pufferfish$(\mathcal{S}, \mathcal{S}_{pairs}, \mathcal{D})$-Privacy if, for all outputs of $\mathcal{M}$, all pairs $(s_i, s_j) \in \mathcal{S}_{pairs}$ and all distributions $d \in \mathcal{D}$ the following holds:

$$P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_i, d) \leq e^\epsilon \cdot P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_j, d)$$
$$P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_j, d) \leq e^\epsilon \cdot P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_i, d)$$

$P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_j, d)$ is the probability that the output of $\mathcal{M}_\epsilon$ is o if s_j holds, and the data distribution is d. □

At first glance, this definition seems complicated. However, we take the equations from Definition 22 and directly compute a more intuitive representation:

$$e^{-\epsilon} \leq \frac{P(s_i | \mathcal{M}_\epsilon(\mathcal{D}ata) = o, d)}{P(s_j | \mathcal{M}_\epsilon(\mathcal{D}ata) = o, d)} / \frac{P(s_i | d)}{P(s_j | d)} \leq e^\epsilon$$

We compare the knowledge of an adversary before (apriori) and after (aposteriori) investigating the data set to gain information. If the adversary apriori thinks s_i is α times as likely as s_j, then after having access to the privacy-enhanced data set $\mathcal{M}_\epsilon(\mathcal{D}ata)$, the adversary aposteriori only believes that s_i is at most $e^\epsilon\alpha$ and at least $e^{-\epsilon}\alpha$ as likely as s_j. Note that Pufferfish [60] itself does not require a specific perturbation method, as long as the guarantees of Definition 22 are fulfilled.

3.1.4 Wavelet transformation

As already explained in the beginning of Chapter 3, it is challenging to specify a representation of private information, which we call secrets, in smart-meter data. Throughout the chapter we will use the well-known wavelet transformation as a sample representation for secrets. Additionally, we explain further transformation methods in Section 3.3.

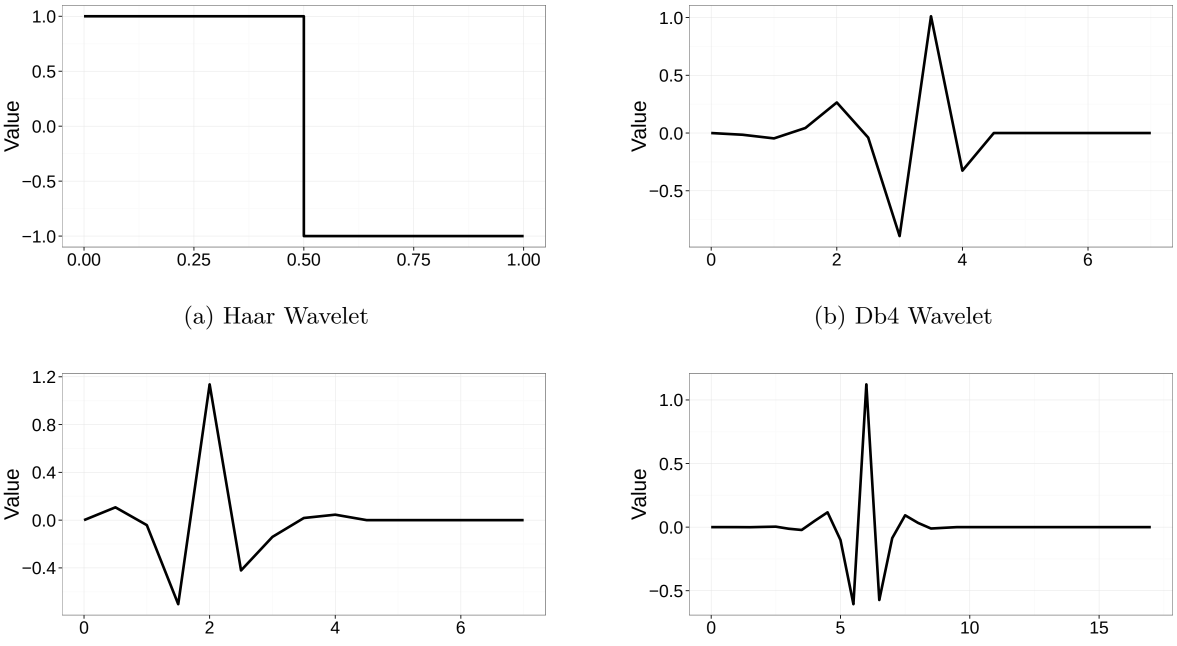
Definition 23 (Wavelet): A wavelet $w[t]$ is a finite time series with the following properties: $\int_{-\infty}^{+\infty} w[t] = 0$ and $\int_{-\infty}^{+\infty} w[t]^2 = 1$. □

Definition 24 (Wavelet Transformation): A wavelet transformation is an orthonormal basis transform to a wavelet basis. Each element of the wavelet basis is a development over time. □

Figure 3.1 contains four sample wavelets that fulfill Definition 23 in the time domain. For our part, we focus on the so-called Haar wavelet (Fig. 3.1a). for the explanation of the wavelet transformation. Definition 23 holds because the area under the curve is of the same size as the area above.

In general, the wavelet transformation is often used for time-series processing. According to Notation 6, the time series is intrepreted as n-dimensional vector with 'time' as a basis. Each basis element represents a time slot, and each vector entry $f[t]$ is the power consumption at time slot t. Wavelet transformation constructs an orthonormal basis, consisting of vectors of time shifted and stretched wavelets. Transforming the time series means changing the basis. Let h be the Haar-wavelet basis for vector f. Then, each element in $f_h[x]$ is relative to the Haar wavelet. The form of the Haar wavelet (Fig. 3.1a) indicates that elements in $f_h[x]$ represent 'changes' in consecutive points of time. In other words, $f_h[x]$ represents the 'Haar pattern'. This intuitive explanation leaves the fact aside that the wavelet as is does not cover the whole vector space because it is naturally considered to be 'short'. Covering the vector space is necessary to provide invertibility. To do so, the wavelet transformation results in multiple levels. Each level corresponds to a horizontally stretched version of the wavelet, and within each level the wavelet is time shifted. The number of levels depends on the length (i.e., dimensionality) of the time series.

As explained previously, the first level always represents the wavelet 'as is'. The higher the level, the more stretched the wavelet itself becomes. For instance, in the second level, a representation with a Haar basis represents the change between, e.g., $f[t], f[t+1]$ and $f[t+2], f[t+3]$. The last level is responsible for the absolute level (in y-dimension) of the time series and does not correspond to the wavelet itself. In the case of the Haar wavelet, the last level does not correspond to any change. In signal processing terms, the lower levels contain the high frequencies, and the higher levels the low frequencies. In contrast to other transformation mechanisms, such as the Fourier



(a) Haar Wavelet

(b) Db4 Wavelet

(c) Sym4 Wavelet

(d) Coif3 Wavelet

Figure 3.1: Sample wavelet functions

transformation [117], each coefficient of the wavelet-transformed representation corresponds to a fixed number of coefficients in the time-based representation. The wavelet transformation keeps the time location.

PACTS is not limited to the wavelet transformation. We define properties that a transformation has to fulfill for its application in PACTS in Section 3.2.1, and show that several popular transformations can be applied (Section 3.3). For a better understanding, we extend the flow heater example to Haar wavelet coefficients.

Example 5 ***(Haar wavelet transformation and flow heaters)*****:** Reconsider Example 4. The starting flow heater leads to a sudden increase in the power consumption. Figure 3.2 illustrates the time-based consumption of a flow heater with the sudden increases. Transforming this time series in the Haar wavelet representation (Fig. 3.4) leads to coefficients smaller than zero for an increase in power consumption and coefficients greater than zero for a decrease in power consumption. Depending on the actual position of the increase or the decrease, this influences coefficients of level one or two. Relevant for hiding Bob's secret are coefficients in f_h, reflecting an increase or decrease of 25 kW. The wavelet representation allows a distinction of whether the flow heater is switched on or off. Thus, it is a suitable way to represent the information that Bob wants to

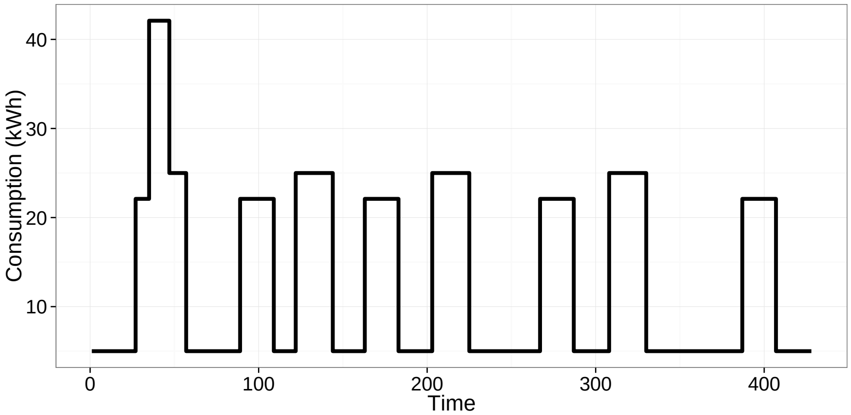


Figure 3.2: Flow heater power consumption in a time-based representation

hide. Depending on the wavelet coefficient level, different values can be relevant (see Fig. 3.3 for an illustration). Although the Haar wavelet represents switch-on or switch-off events well, other wavelets (see Fig. 3.1) represent different patterns. In conclusion, a similar approach is possible for different discriminative pairs. □

3.2 PACTS: Provable Privacy for Smart-Meter Time Series

In this section, we explain PACTS, an instantiation of the Pufferfish privacy mechanism $\mathcal{M}_\epsilon$ for smart-meter data. $\mathcal{M}_\epsilon(f)$ transfers time-series f into one that guarantees ϵ-Pufferfish privacy with respect to the given parameters. More precisely, $\mathcal{M}_\epsilon$ conducts the steps illustrated in Figure 3.5. To ease the presentation, we assume a single discriminative pair of secrets s_{pair} in the following. This does not restrict the generality of our approach, because each $s_{pair} \in \mathcal{S}_{pairs}$ is handled in isolation. For further explanation, see Algorithm 3, which contains the pseudo-code implementation of PACTS for arbitrary sets of discriminative pairs and time series. Specifically, we explain in the following steps the loop body of Algorithm 3.

Step 1. In the first step, we transform the time-series f into an abstracted representation f_B. We use such representations to isolate information on a user-defined discriminative pair to single coefficients. Reconsider Example 4: In the Haar wavelet–transformed representation, certain single coefficients determine whether the flow heater starts or stops. We elaborate more on the technical details in Section 3.2.1.

Step 2. In the abstracted representation f_B, we perturb the time series according to user-defined preferences. The resulting abstracted time series guarantees Pufferfish privacy. Among others, this

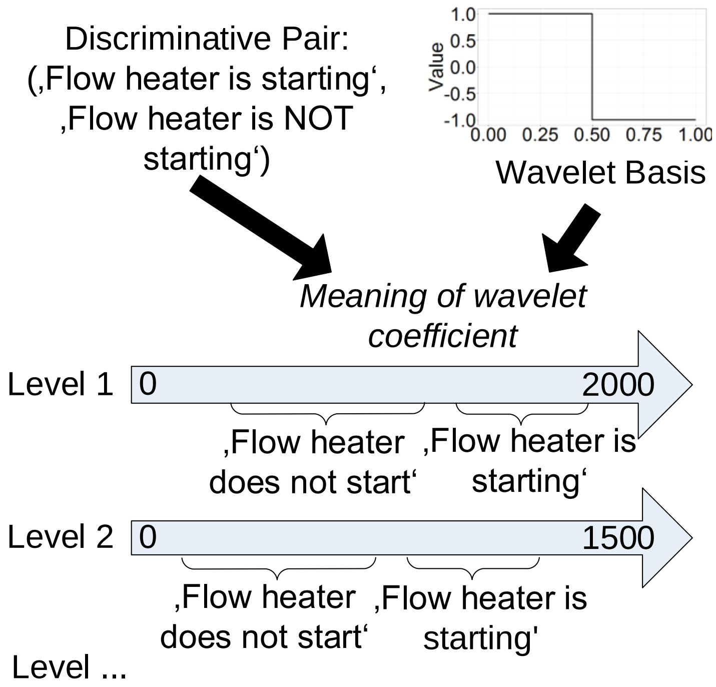
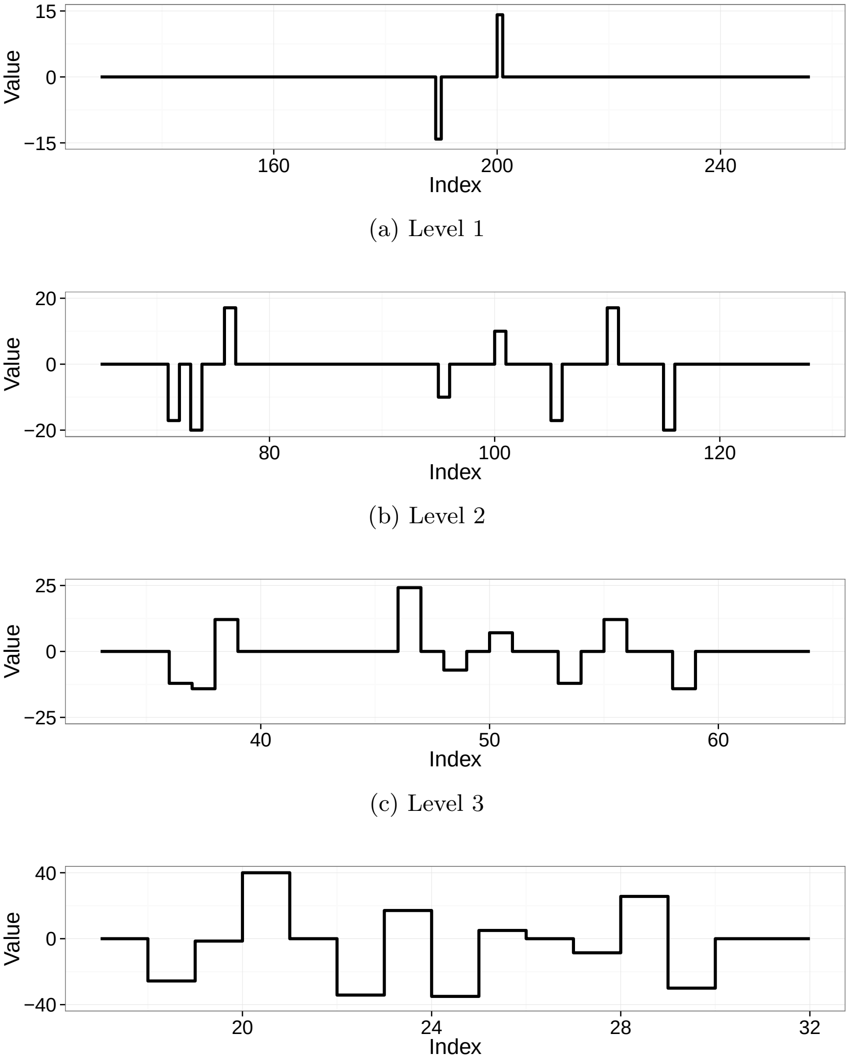

Discriminative Pair:
(‚Flow heater is starting‘,
‚Flow heater is NOT
starting‘)
1.0
0.5
0.0
-0.5
-1.0
Value
0.00 0.25 0.50 0.75 1.00
Wavelet Basis
Meaning of wavelet coefficient
Level 1
0
2000
‚Flow heater does not start‘
‚Flow heater is starting‘
Level 2
0
1500
‚Flow heater does not start‘
‚Flow heater is starting'
Level ...

Figure 3.3: Sample mapping of Haar wavelet coefficients to a starting/stopping flow heater



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

Figure 3.4: Haar wavelet–transformed representation of the flow heater power signal

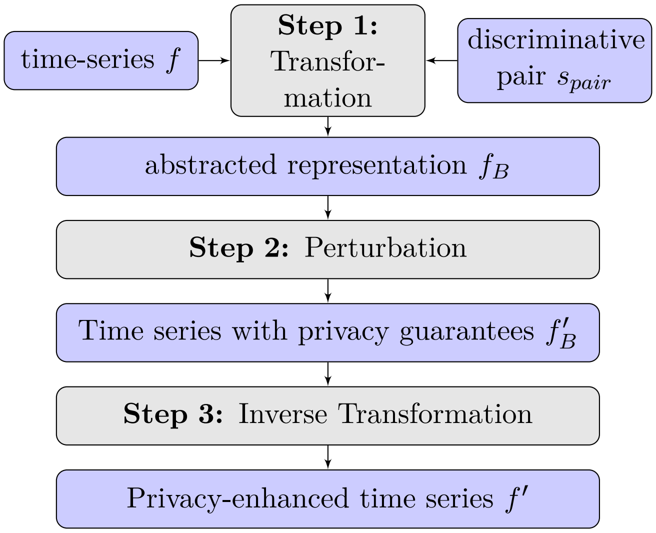

time-series f
Step 1: Transformation
discriminative pair s_{pair}
abstracted representation f_B
Step 2: Perturbation
Time series with privacy guarantees f'_B
Step 3: Inverse Transformation
Privacy-enhanced time series f'

Figure 3.5: Privacy preserving for a single discriminative pair s_{pair}

step requires the determination of the noise distribution. Section 3.2.2 explains the details.

Step 3. In the last step, we transform the (privacy-enhanced) time series back to the time-based representation f' (Section 3.2.3).

3.2.1 Step 1: Transformation

In this step we transform a given time-series f to an abstracted representation f_B. Each coefficient $f_B[t]$ carries a specific meaning for secrets and not necessarily to a point of time anymore. Secrets require a transformation to specify relevant coefficients. Thus, we define and specify the transformation mechanism before formulating secrets and discriminative pairs for smart-meter data.

Transformation Mechanism

Representations of time series in an abstracted manner are numerous. Fourier and wavelet transformations are well-known, but there are several others [23]. The right choice depends on the required discriminative pair. To keep the approach general, we define properties that each transformation has to fulfill to be applicable in PACTS.

Definition 25 (Transformation Mechanism): Let E be the standard basis and B a different basis of vector space $\mathcal{V}_B$. A transformation mechanism $\mathcal{C}_B$ is a function of type $\mathcal{V}_E \rightarrow \mathcal{V}_B$ that converts a time series from the standard basis f_E to an abstracted representation f_B and fulfills the following properties:

Algorithm 3: PACTS-Pufferfish Privacy Mechanism $\mathcal{M}_\epsilon$ for Smart-Meter Data

```
Input: time series $f$
Input: Set of discriminative pairs $\mathcal{S}_{pairs}$ of secrets $\mathcal{S}$, (Inverse) Transformation Mechanism
       $\mathcal{C}_B^{trans}$, $\mathcal{IC}_B$ and basis $B$
Input: Data evolution scenarios $\mathcal{D}$
Input: Privacy parameter $\epsilon$
Result: Time series with privacy guarantees $f'$
foreach $s_{pair} \in \mathcal{S}_{pairs}$ do
    // Step 1: Transformation;
    $f_B = \mathcal{C}_B^{trans}(f)$;
    // Step 2: Perturbation;
    Determine $\mathcal{N}_\epsilon$ to fulfill $\epsilon$-Pufferfish Privacy based on $\mathcal{D}$ and $s_{pair}$;
    Set $p^{coeff}$ according to $s_{pair}$;
    $f'_B = \mathcal{P}(f_B, \mathcal{N}_\epsilon, p^{coeff})$;
    // Step 3: Inverse Transformation;
    $f' = \mathcal{IC}_B(f'_B)$
end
return $f'$;
```

1. The transformation is invertible, i.e., there exists an inverse of $\mathcal{C}_B$ that we refer to it as $\mathcal{IC}_E$. We precisely define the inverse in Section 3.2.3; and

2. $\mathcal{C}_B$ has to be an endomorphism for the +-operator. Let f, g be the time series defined on the same basis E, then $\mathcal{C}_B(f + g) = \mathcal{C}_B(f) + \mathcal{C}_B(g)$.

□

The endomorphism is an important property because of the following reason. Assume that a time series is an aggregate of different power consumptions. If a smart meter records time series, this is the natural case, because all appliances are connected together to the main power supply. The endomorphism property simplifies the perturbation conducted in Step 2 (Section 3.2.2). Noise can be added to the aggregate and to certain parts of the aggregate in the exact same way.

The invertibility property implies the following: An abstracted representation of a time series f_B, which is invertible, obviously has to contain all the information present in f. In other words, no information is lost during transformation. Additionally, invertibility requires well-defined semantics of every element in f_B. Those clear semantics also hold for the definition of secrets, i.e., each coefficient has a specific meaning in relation to a secret.

Definition 25 does not restrict the dimensionality of the transformed time series. The transformation output f_B might have a higher or a lower dimensionality compared with f_E. Usually, the dimensionality of f_B will be at least as high as the dimensionality of the nontransformed representation to ensure invertibility.

Haar-Wavelet Transformation for PACTS The wavelet transformation introduced in Section 3.1.4 satisfies Definition 25. In particular for the 'Haar' wavelet, we show in Lemma 1 that the transformation is invertible and an endomorphism for addition. Thus, it is applicable in PACTS. Additionally, the wavelet transformation keeps the time location. In other words, each coefficient $f_h[x]$ corresponds to specific entries in $f_h[t]$.
Notation 9 (Haar Wavelet Transformation $\mathcal{C}_h^{Wave}$): $\mathcal{C}_h^{Wave}$ denotes the wavelet transformation with the Haar basis in PACTS.

Lemma 1: *The Haar wavelet transformation is invertible and an endomorphism for the +-operator.*
Proof: The Haar wavelet transformation defines an orthonormal basis for any vector with $2^n, n \in \mathcal{N}$ coefficients [24]. Thus, this basis forms an orthonormal basis transformation matrix H that changes the basis of a vector as follows:

$$f \cdot H = f_h$$

For each orthonormal matrix, an inverse H' exists that is also orthonormal. Letting I be the identity matrix, the following holds: $H \cdot H' = I$. With the help of H', it is straightforward to show invertibility:

$$f_h \cdot H' = f \cdot H \cdot H' = f \cdot I = f$$

It is well known that the matrix vector multiplication is distributive:

$$f \cdot H = (f^1 + \dots + f^i) \cdot H = f^1 \cdot H + \dots + f^i \cdot H$$

Thus, the Haar wavelet transformation is also an endomorphism for the +-operator. □

Secrets in Smart-Meter Data

Secrets $\mathcal{S}$ are the core of user-defined-privacy requirements. They express the information deemed sensitive. Such requirements range from relatively simple ones like *'The dishwasher is running'* to rather complex ones involving several appliances like *'There is cooking activity'*. Other examples might be *'There is activity in the kitchen'*, *'The fridge is running'* or *'Someone is watching a certain TV program in the morning'*. A secret involving several appliances is more complex because their power consumptions may overlap differently each time the secret is true. Additionally, secrets may also be relevant only for certain periods, i.e., an individual deems an activity only during a certain time sensitive.

The power-consumption data of individual households are usually monitored by a smart meter installed at the main power connection. Thus, the smart-meter data are an aggregate of all appliances. However, following the intuition of the presented examples, only parts of this aggregate are typically relevant for a secret. Hence, it is important to be able to examine parts of the smart-meter time series in isolation. Regarding the time series as a signal, it is the aggregate of several channels. For example, the consumption of the television can be considered as one channel $f^1[t]$, whereas the dishwasher is another one $f^2[t]$.

Definition 26 (Signals and Channels): A signal is the total power consumption measured at the smart meter of the household, and is represented as vector $f[t]$. A channel is part of the mentioned signal, referred to with a superscript, e.g., $f^i[t]$. Consequently, a signal is the sum of n channels.

$$f[t] = f^1[t] + \cdots + f^n[t]$$

□

The channels isolate relevant appliances from others. Still, a sequence of consumption values is required in many cases to gain information. The fact that a sequence of time-value pairs identifies appliances and their state is well known from non-intrusive appliance load monitoring (NIALM) approaches [42, 45, 68, 83, 121], and appliances tend to be detectable in the signal.

Even in the abstracted representation, the relation between the intuitive secret description and coefficients is not straightforward. The following definition allows us to specify these relations.

Definition 27 (Description of a Secret): A description of a secret is a triple

$$s = (s^{Base}, s^{Trans}, s^{Coeff})$$

where s^{Base} is the basis for the transformation mechanism s^{Trans} according to Definition 25. s^{Coeff} is the formal description of the coefficients in the abstracted representation f_B ($s^{Base} = B$), which make s true. $f_B[t] \in s^{Coeff}$ denotes that coefficient t of the transformed time series makes the secret true. □

The definition does not require a formal specification of the language describing s^{Coeff}. However, the description has to be nonambiguous.

The description of secret s only reflects *what* should be hidden, but not *how*. To ensure Pufferfish privacy, PACTS requires discriminative pairs of secrets describing the *'how'*. Intuitively, it makes a difference to hide which kind of appliances are running compared with hiding different running states. Hiding the running states usually tends to require less noise instead of hiding the kind of appliance running. The following definition formalizes the description of discriminative pairs in PACTS.

Definition 28 (Description of a Discriminative Pair of Secrets): A description of a discriminative pair of secrets s_{pair} is a pair of description of secrets $s_{pair} = (s_1, s_2)$, so that the following holds:

- The base and the transformation method are the same, i.e., $s_1^{Base} = s_2^{Base}$ and $s_1^{Trans} = s_2^{Trans}$;

- The secrets do not need to be exhaustive, i.e., there may exist values in the range of a coefficient that neither makes s_1 nor s_2 true; and
- The secrets are mutually exclusive, i.e., at least one is true. Thus, the coefficients in question for s_1 and s_2 are non-overlapping: $s_1^{Coeff} \cap s_2^{Coeff} = \varnothing$.

□

To distinguish whether a secret of a discriminative pair is true, typically only parts of the entire signals are relevant. We explicitly call those channels 'relevant'.

Definition 29 (Relevant Channel): For a given discriminative pair $s_{pair} = (s_1, s_2)$, we call the channel that contains all of the relevant information, whether s_1,s_2 or none is true about the relevant channel r. If the signal f consists of $i \in [1 \dots n]$ channels, we refer to the relevant channel as r and to the corresponding vector as f^r. According to Definition 26, the decomposition partitions the whole signal:

$$f[t] = f^1[t] + \dots + f^r[t] + \dots + f^n[t]$$

□

The data contained in different channels may be statistically independent or not. The following example shows that this depends on the considered discriminative pair and the assumptions of an adversary. According to the Pufferfish privacy framework, the required distribution of noise depends on these assumptions. Assumptions regarding statistical distributions and dependence result in data-evolution scenarios. We will elaborate more on these assumptions in Section 3.2.2.

Example 6 ***(Statistically Dependent and Independent Channels)*****:** Assume that channel f^1 contains the TV only. Thus, typically f^1 is uncorrelated with channel f^2, which contains the dishwasher. If the discriminative pair in question only refers to the TV, and the running dishwasher does not influence the TV, both channels are statistically independent. In turn, think of a discriminative pair containing the secrets 'The household is cooking.' and 'The household is not cooking'. In this case, there are most likely correlations among several devices in the kitchen such as appliances like the oven and the kitchen lighting. However, the lighting is not part of the relevant channel, as cooking does not directly relate to light in the kitchen. However, kitchen lighting indicates activity in that room and is not statistically independent of the appliances. □

Reconsider Example 2: After having defined secrets and discriminative pairs in smart-meter data, we can provide proper descriptions for PACTS. The following example illustrates the description of secrets and of the corresponding discriminative pair.

Example 7 ***(Transformation and Instantiations of Secrets for the Flow Heater)*****:** Bob wants to hide whether the secret s_1 'The flow heater is starting/stopping' or secret s_2 'The flow heater is not starting/stopping' is true. We have already seen that the wavelet transformation with the Haar basis reflects 'switch on', respectively 'switch off' events, well (see Example 4), and is a suitable transformation for the discriminative pair $s_{pair} = (s_1, s_2)$ in PACTS. For both secrets we choose the Haar-wavelet transformation: $s_1^{Trans} = s_2^{Trans} = \mathcal{C}_h^{Wave}$. For the sake of simplicity, we assume that the flow-heater power consumption is a rectangular shape over time, as illustrated in Figure 3.2. We generated the consumption with the model of [100]. Figure 3.4 contains $\mathcal{C}_h^{Wave}(f)$ of the time series illustrated in Figure 3.2: The x-axis in the transformed representation (Fig. 3.4) shows the time location and the y-axis the 'intensity' of the Haar basis. Coefficients in Levels 1 and 2 reflect the starting and stopping of the flow heater according to the explanation in Section 3.1.4. To include small inaccuracies, we define s_1^{Coeff} to contain coefficients of Level 1 if their value is $[13, 17]$ or $[-17, -13]$, and Level 2 if their value is $[18, 22]$ or $[-22, -18]$. In turn, s_2^{Coeff} contains all values of coefficients on Level 1 except $[13, 17]$ and $[-17, -13]$ and Level 2 except for $[18, 22]$ or $[-22, -18]$. Obviously, $s_1^{Coeff} \cap s_2^{Coeff} = \varnothing$ and $s_1^{Trans} = s_2^{Trans}$. Thus, $s_{pair} = (s_1, s_2)$ qualifies as a description of a discriminative pair according to Definition 28. Furthermore, in this example the relevant channel contains only the flow-heater consumption values. □

For different bases or transformations, the determination of the coefficients works similarly. Different ways of transforming time series cover different privacy requirements. We will discuss further transformations in Section 3.3.

3.2.2 Step 2: Perturbation

In this section, we explain how PACTS ensures the ϵ-Pufferfish privacy principle in the time series of smart-meter data. One common method is to apply additive noise following the Laplace distribution to aggregates [60]. As explained previously in Section 3.2.1, the smart-meter signal is the aggregate of the household's appliances. Discriminative pairs are usually limited to the relevant channel. Thus, noise is only required for some channels. Identifying these channels, and the distribution of the noise applied, is not obvious. In this section we explain the necessary steps in PACTS to apply noise.

Perturbation Mechanism for Time Series

We explain how PACTS perturbs the time series of smart-meter data in the transformed representation. Naturally, this requires a noise distribution $\mathcal{N}_\epsilon$. We refer to the discriminative pair s_{pair} with both secrets having the same basis $s^{Base} = B$ and the same transformation mechanism $s^{Trans} = \mathcal{C}_B$. We refer to the perturbed time series as f'_B. In addition, the perturbation also requires the selection of coefficients to be noised. This leads to the following definition.

Definition 30 (Perturbation Mechanism for a Discriminative Pair): A perturbation mechanism $\mathcal{P}$ is a function that takes a time series f_B in the abstracted representation, the applied noise $\mathcal{N}_\epsilon$ that is dependent on the privacy parameter ϵ and a formal definition of the coefficients to be perturbed p^{coeff}. It returns the privacy-enhanced time series in the abstracted representation, referred to as f'_B.

$$f'_B = \mathcal{P}(f_B, \mathcal{N}_\epsilon, p^{coeff})$$

□

Noised Elements

p^{coeff} specifies the elements of the abstracted time series f_B to be perturbed. As in the definition of the secret description, we do not require a formal language for selecting these coefficients. However, we provide a classification. Depending on the actual transformation mechanism, not all kinds of coefficients are possible.

- **All:** This is the most simple strategy. Additive noise is applied to all coefficients.
- **Trigger dependent:** As already seen, coefficients in a certain range have a defined meaning. A discriminative pair might require the addition of noise only if a coefficient corresponds to a certain meaning.
- **Time dependent:** The user specifies coefficients to be perturbed if they correspond to a certain time interval in the time-based representation. In comparison to the trigger-dependent approach, the value of the coefficient does not play a role. However, this only works if the transformation mechanism keeps the time location.
- **Trigger and time dependent:** The combination of both is obviously also possible.

Noise Distribution

$\mathcal{P}$ used with noise according to the Pufferfish privacy principle and to the discriminative pair $s_{pair} = (s_1, s_2)$ will guarantee privacy. The following lemma proves this for PACTS.

Lemma 2: *Let f be a time series of smart-meter data, $s_{pair} = (s_1, s_2)$ the information that an individual wants to hide, $\mathcal{C}_B$ the transformation mechanism suitable for s_{pair}, and $\mathcal{P}$ a perturbation mechanism. A distribution of noise $\mathcal{N}_\epsilon$ exists for $\mathcal{P}$ such that $f'_B = \mathcal{P}(f_B, \mathcal{N}_\epsilon, p^{coeff})$ satisfies the ϵ-Pufferfish privacy definition.*

Proof: In PACTS, secrets (Definition 27) and discriminative pairs (Definition 28) are defined according to the Pufferfish framework [60]. Assume that the data evolution scenario $\mathcal{D}$ defines the distribution of values on each channel of the whole signal f_B, including the relevant channel r for s_{pair} f_B^r. Because the transformation mechanism is an endomorphism for the +-operator, applying additive noise (even to single channels) is possible regardless of the actual representation and even if only the signal (and not the whole decomposition) is available. If we apply noise $\mathcal{N}_\epsilon$ for $s_{pair} = (s_1, s_2)$ so that the following holds, ϵ-Pufferfish privacy is guaranteed:

$$P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_1, s_{pair}) \leq e^\epsilon \cdot P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_2, s_{pair})$$
$$P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_2, s_{pair}) \leq e^\epsilon \cdot P(\mathcal{M}_\epsilon(\mathcal{D}ata) = o | s_1, s_{pair})$$

According to [60], a suitable distribution of additive noise can be found for every fixed $\mathcal{D}$ dependent on ϵ. □

The following example illustrates how to choose appropriate noise for the starting flow heater.

Example 8 *(Hiding the start of the flow heater)*: Reconsider Example 7: Bob wants to hide the the discriminative pair $s_{pair} = (s_1, s_2)$ where s_1 = 'Flow heater is starting' and s_2 = 'Flow heater is NOT starting'. Note: This example covers the case when the flow heater is starting. Hiding the stop of a flow heater is similar, i.e., the wavelet coefficient for a switch-off event is the same as for the switch-on events with an inverted sign. To do so, we carry out the transformation from Example 5 with the wavelet transformation $\mathcal{C}_h^{Wave}$ and the Haar basis h. Let f^r be the relevant channel for s_{pair}. To ease the presentation, suppose that the channels are statistically independent. According to Example 7, the coefficients in question for s_1 or s_2 correspond to nonoverlapping intervals by definition. For instance, let $f_h[x]$ be a value of Level 1 of the wavelet-transformed representation. If $f_h^r[t] \in [y-k, y+k]$, s_1 is true for $y = 15$ with an imprecision interval of $k = 2$, otherwise s_2. For Level 2 s_1 is true for $y = 20$ and $k = 2$. Bob wants to prevent an adversary from learning the value of $f_h^r[x]$ by accessing the privacy-enhanced signal f_h'. [60] shows that adding noise drawn from the Laplace$(4k/\epsilon)$ distribution with density function $\frac{\epsilon}{8k}e^{-\epsilon|x|/4k}$ guarantees ϵ-Pufferfish privacy for an aggregate as follows: An adversary cannot distinguish whether the value of a single channel is between $y-k$ and $y+k$ or one of the neighboring intervals $[y+k, y+3k)$ or $[y-3k, y-k)$. Let X be a random variable drawn from the Laplace$(4k/\epsilon)$ distribution and t the coefficent to hide. We then generate the privacy-enhanced aggregate $f_h'[x]$ as follows:

$$f_h'[x] = f_h^r[x] + f_h^i[x] + \dots + X$$

□

In this case, adding noise does not require the decomposition of the signal into several channels, i.e., $f_h'[x] = f_h[x] + X$. Adding noise to the signal already ensures Pufferfish privacy; however, the definition of the distribution requires knowledge about the relevant channel, respectively, about consumption patterns that are present there.

With the wavelet transformation, time location is also possible. Following Example 2 we add noise for weekdays between 8:00 and 11:00 only. On the weekends, we add the noise defined for the whole day.

3.2.3 Step 3: Inverse Transformation

Before disclosure, the last step transforms the abstracted and perturbed representation f_B' back to the time-based representation f'.

Definition 31 (Inverse Transformation): An inverse transformation mechanism $\mathcal{IC}_B$ is a function that takes a time series in abstracted representation f_B and returns the same time series in the time-based representation f. □

Because Definition 25 requires invertibility, an inverse transformation mechanism $\mathcal{IC}_B$ exists for each $\mathcal{C}_B$.

3.3 Transformations

Until this point we have introduced PACTS as a way of guaranteeing ϵ-Pufferfish privacy on the time series. However, there are still issues worth discussing. First, we have illustrated in several examples that the Haar wavelet transformation is applicable to hide the switch on and off events of a flow heater. Because the Haar wavelet reflects (sudden) increases and decreases of the power consumption, this transformation is also applicable for other single-switching events. We discuss expressiveness and limitations of the Haar-wavelet transformation in Section 3.3.2. Second, other secrets may require completely different transformations. We discuss alternatives to the Haar-wavelet transformation for PACTS in Section 3.3.2.

3.3.1 Expressiveness of the Wavelet Transformation

We have already introduced the so-called non-intrusive-appliance-load monitoring (NIALM) approaches in Section 2.1.1. Their common goal is to extract information about the running state of appliances by inspecting the aggregated power consumption. During the detection process, a major role is the switch on and off events that can be seen in the aggregated consumption as sudden increases. Running appliances corresponds to different specific activities in the household. Thus, it is promising to completely hide these events to protect the privacy of individuals. Furthermore, one feature for re-identification is the Wakeup (Definition 13), respectively the Bedtime-hour (Definition 14), also corresponding to a sudden increase or decrease in the power consumption. In the evaluation (Section 3.4), we will define the secrets that hinder the INDiC NIALM approach to detect appliances, in addition to the mentioned features of re-identification.

However, the Haar-wavelet transformation also has two limitations: First, the Haar basis is of length two and can therefore only transform time series of length 2^n. Second, it is not trivial to find another basis that describes other patterns. Remember that each wavelet has to fulfill Definition 23. Modified wavelet transformations and completely different approaches are introduced in the next section.

3.3.2 Transformation Mechanism

If a transformation fulfills Definition 25, it can be used in PACTS to hide discriminative pairs of secrets. One promising way is to take the exact same transformations that an adversary will use to extract information on the discriminative pair. For instance, such approaches are part of the NIALM approaches (see Section 2.1.1). However, this usually requires additional effort to ensure the required properties for a transformation mechanism in PACTS. There are numerous other well-known transformations that could be used instead of the presented ones without modifications. We will see in the evaluation (Section 3.4) that the provided transformations are general enough to cover a wide range of secrets.

Decomposed Wavelet Transformation

The wavelet transformation is capable of transforming a time series if its length is 2^n. In general, this is not the case, but we can decompose the signal: The decomposed wavelet transformation splits the original signal into different disjoint sub-sequences and applies the wavelet transformation to each one. This allows independent modification of different periods of the signal. A popular decomposition is the ancient Egyptian decomposition [19]. This decomposition finds the largest power of two less than or equal to the length of the time series in question. This forms the first sub-sequence and the total length is reduced by this number. This step is repeated until the whole signal is decomposed into sequences, each with a length of the power of two.

Lemma 3: *The decomposed wavelet transformation fulfills Definition 25, i.e., is invertible and an endomorphism for the +-operator.*

Proof: In Lemma 1 we have already shown that a wavelet transformation fulfills the necessary requirements. The decomposed transformation processes distinct parts of the time series and is also invertible and an endomorphism. □

Wavelet-Packet Transformation

The wavelet-packet transform is another wavelet transformation. In contrast to the transformation already proposed, it does not require a specified basis such as the Haar basis. In particular, with the help of a time series representing the pattern of a secret, the packet transform is able to compute a suitable basis. The resulting basis is matched to the given time series [21]. The advantage of the packet transform is that it can be used to flexibly create wavelet bases that match patterns well. Such a precomputed basis is used to transform the signal, respectively the channels, following the standard wavelet transformation. Although the wavelet-packet transformation provides further flexibility, we do not use it in our evaluation in Section 3.4, as other transformations suffice to deal with the secrets featured there.

Lemma 4: *The wavelet-packet transformation fulfills Definition 25, i.e., is invertible and an endomorphism for the +-operator.*

Proof: The wavelet packet transformation chooses a custom base for the transformation as a composition of orthonormal bases. Thus, it is invertible. Because the transformation applies the same basis to all of the channels, the addition of the coefficients is well-defined and thus is also an endomorphism for the + operator. □

Discrete Fourier Transformation

Oscillations in the power consumption are periodically repeating power demands, e.g., appliances running at fixed times. Oscillations are also a characteristic of the state of appliances, e.g., the frequency of power peaks of a television corresponds to the TV program. The discrete Fourier transformation (DFT) [92] converts a sequence of samples (this is the time series) to a frequency-decomposed representation of the described oscillations. Thus, this transformation allows one to hide periodical events.

Lemma 5: *The discrete Fourier transformation fulfills Definition 25, i.e., is invertible and an endomorphism for the +-operator.*

Proof: Each coefficient in the Fourier-transformed representation corresponds to certain well-defined frequencies. Thus, there exists an inverse transformation [117]. Furthermore, the value of each coefficient is the amplitude of a certain frequency. A sum in the time domain of two time series equals the sum of all frequency amplitudes. The DFT is also an endomorphism for +, and conclusively fulfills Definition 25. □

Codebooks and Multiresolution Analysis

Individuals might have a certain pattern in mind that they want to hide and then use a multiresolution-codebook representation such as [114] to search for this pattern. In a nutshell, a codebook is a map from keys to patterns (sequences of power-consumption values). The abstracted time series is represented by a sequence of these keys, and each value corresponds to the pattern described by code words in the codebook. In general, there may be a small difference between the codewords and the actual patterns. Usually these differences are neglected [114], leading to an inaccurate inverse. Invertibility requires recording these differences. Patterns can also be created by compression algorithms [18, 119] such as LZW, which extract similar sequences. Whether such transformations fulfill the requirements of Definition 25 depends on the actual algorithm. A codebook is invertible because it is a unique map. It is also an endomorphism for + if the addition of two keys results in a key that represents the addition of the patterns in the time domain.

3.4 Evaluation

Evaluating PACTS has two main goals: generality and utility. First, an individual should be able to hide arbitrary information as discriminative pairs. Second, the privacy-enhanced data should still be useful even while guaranteeing privacy to the extent specified.

Regarding generality, to evaluate in an objective way whether PACTS is general enough to cover a broad range of privacy requirements, we need a reliable and objective source for such requirements. However, such a source is difficult to find. Studies like [36] indicate that statements of individuals regarding privacy are rather contradictory. To the best of our knowledge, a source containing individually defined privacy requirements for smart-meter data does not exist. However, recent

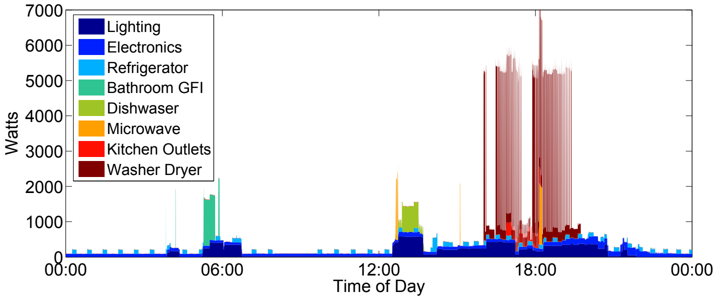


Figure 3.6: REDD data disaggregation example [62]

approaches have extracted various kinds of information about individuals from smart-meter data. In turn, the information that these approaches try to extract can be perceived as information that is worth protecting, i.e., formulated as privacy requirements. We show that it is possible to define discriminative pairs that are suitable for these requirements. The approaches explicitly considered are INDiC [8], a NIALM approach, and the re-identification introduced in Section 2.2. In total, we have identified over 13 categories of secrets. We will show that guaranteeing Pufferfish privacy with PACTS makes information extraction with these methods significantly more difficult. The NIALM approach requires data from smart meter as well as from individual devices. We chose the REDD data set as described in Section 3.4.1.

The issue of utility is the second important evaluation dimension. Abstract time-series-distance measures do not allow for meaningful conclusions regarding the utility of a privacy-enhanced time series. We elaborate on this and introduce an application-specific measure in Chapter 4 by means of a local electricity market.

3.4.1 Reference Energy Disaggregation Data Set (REDD)

The Reference Energy Disaggregation Data Set (REDD) [62] is a publicly available smart-meter data set. Its main purpose is to provide a benchmark in the field of energy disaggregation, which is the task of determining which and to what extent appliances influence the measured aggregate electricity signal. The evaluation of disaggregation methods against REDD allows the comparison of such methods. To have a ground truth, the data set contains the electric power consumption that is measured at the main power supply (aggregated) and the consumption at a number of outlets. Figure 3.6 illustrates a sample disaggregation.

In particular, smart meters were installed in six houses. The installation included the main power connection, which was divided into two channels, and up to 24 outlets were mapped to specific appliances. The smart meter recorded the data for several months. REDD contains 'low frequency' measurements every 1 or 3 s and 'high frequency' measurements at 15 kHz. In this work, only the measurements with low frequency are of relevance.

3.4.2 Generality: The INDiC NIALM Approach

As a first step in evaluating generality, we assume that individuals want to hide whether a specific appliance is running. NIALM approaches allow the extraction of running appliances from the aggregated smart-meter signal. Although the different NIALM methods are numerous, we choose INDiC [8], a refinement of one of the first methods [45]. For a detailed explanation, see Section 2.1.1.

Evaluating how well secrets hinder information extraction with INDiC requires a ground truth. It contains whether INDiC is successful when extracting information on running devices. Thus, the creation of the ground truth requires the smart-meter signal and individual channels of devices to compute success rates. We use the publicly available REDD data set [62], which contains the total power consumption of different households divided into two 'main' signals (smart meter) and a number of isolated channels (electricity outlets) monitored in parallel. A detailed explanation of the data set is provided in the previous Section 3.4.1. The disaggregation, together with the subsequent evaluation, consists of the following steps:

1. The data set (including both main and appliance channels) is divided into a training and a test set;

2. For each appliance channel available, INDiC determines the possible different states by clustering the power-consumption values of the training set;

3. Based on the states identified, the main channels in the test data set are disaggregated; and

4. To evaluate the success of the disaggregation, the computed results are compared with the actual appliance-usage data available from the other channels.

Application of PACTS

When defining the secret descriptions, we require knowledge of devices: Table 3.4 lists the results of the training for all provided appliances. As a result of the training, INDiC comes up with different states of appliances by finding frequent power-consumption levels. Each level corresponds to a specific state, and the number of states may vary contingent on the kind of appliance. The states with the corresponding power level are the external knowledge of an adversary trying to gather information by inspecting the aggregated power consumption time series f. INDiC determines running appliances by attributing the total power consumption to states.

In this part of our evaluation, we assume that an individual wants to keep INDiC from determining the states that appliances are actually in. Without loss of generality, we choose three discriminative pairs, with secrets corresponding to the same appliance being in two different states. In the following we describe the application of the discriminative pair regarding the 'light' appliance. Choosing another pair only requires the use of other power-consumption levels in the secret. We summarize the evaluated pairs and the necessary parameters in Table 3.3.

s_1	s_2	k
‘Light is in State 2’	‘Light is in State 3’	$k = (153\ W - 113\ W)/2 = 20\ W$
‘Refrigerator is in State 2’	‘Refrigerator is in State 3’	$k = (423\ W - 214\ W)/2 = 104.5\ W$
‘Microwave is in State 2’	‘Microwave is in State 3’	$k = (1740\ W - 822\ W)/2 = 459\ W$

Table 3.3: Discriminative pairs $s_{pair} = (s_1, s_2)$ for the INDiC generality evaluation

The intuitive description of the ‘light’ secret is s_1 = ‘Light is in State 2’ and s_2 = ‘Light is in State 3’. INDiC works without modifying the representation of the time series. Hence, we modify the time series as follows: $s_1^{Trans} = s_2^{Trans} = id$, and the base is $s_1^{Base} = s_2^{Base} = E$. According to Table 3.4, light is in State 2 if 113 W is not attributed to another appliance and in State 3 if 156 W is not attributed elsewhere. s_1^{Coeff} contains coefficients that result in 113 W, and s_2^{Coeff} contains coefficients that result in 156 W of unaccounted power. Then the discriminative pair is $s_{pair} = (s_1, s_2)$. INDiC assumes that all appliances have the same probability to be in a specific state, i.e., we can assume that $\mathcal{D}$ is evenly distributed when adding noise. Because the secrets considered do not specify a time span, we set p^{coeff} to f. In summary, an adversary should be unable to distinguish whether the unaccounted power is approximately 113 W or 156 W. According to Example 8, we choose Laplace$(4 \cdot k/\epsilon)$ and $k = \frac{153-113}{2}$ noise perturb the interval between both values. Furthermore, we assume that the individuals require to achieve ϵ-Pufferfish privacy with $\epsilon = 0.1$.

Known Limitation: Note that Laplace noise can also yield negative values. Depending on the noise applied and the actual total power consumption, values lower than zero are possible. If negative values occur, we set them to zero to ensure that the resulting data are valid. However, this clearly influences the noise distribution required to hide discriminative pairs in PACTS. Theoretically, one may not be able to guarantee privacy when large differences between states shall be hidden. However, this is not specific to Pufferfish or to PACTS. Instead, this is a general problem of information-hiding approaches; perturbing information that is a significant part of an aggregated value requires noise with a large variance. The actual effects of this measure can be seen in the results (Section 3.4.2).

Results

To quantify the error resulting from the noise, we conducted an INDiC disaggregation on the test data set with and without noise applied. We determine the loss of accuracy and the change in uncertainty, whether the appliance considered is in a state of secret s_1 or secret s_2. To examine the results of the INDiC disaggregation, we choose confusion matrices as proposed in [8]. The rows represent the predicted state of the appliance, and the columns the actual state determined from ground truth data. Thus, the element at $m \times n$ represents the frequency when the mth state

outlet/appliance	**State 1**	**State 2**	**State 3**
dishwasher	$0\ W$	$260\ W$	$1,195\ W$
kitchen	$5\ W$	$727\ W$	
kitchen2	$1\ W$	$204\ W$	$1,036\ W$
light	$9\ W$	$113\ W$	$156\ W$
microwave	$9\ W$	$822\ W$	$1,740\ W$
refrigerator	$7\ W$	$214\ W$	$423\ W$
stove	$0\ W$	$373\ W$	

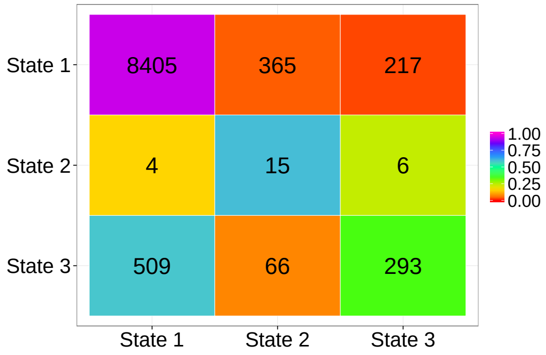
Table 3.4: States of appliances

was detected while the state was actually n. We consider absolute and relative frequencies. After applying noise as described in Section 3.4.2, we expect the results to get worse.

R.11 *The INDiC approach determines the states of appliances well.* For the 'refrigerator' and the 'microwave' appliances, the INDiC determines in more than 75% accuracy (Figs. 3.9a and 3.11a). Assuming that each of the seven appliances with 19 possible states are equally distributed, the INDiC results are significantly better than random guessing. For the 'light' appliance, the results are worse: State 3 is correctly determined only in 46% of the cases, State 1 is correctly determined in 94%, and State 2 in 60%. Because the difference in the consumption levels of 'light' between State 2 and State 3 are relatively low (c.f., Table 3.4), it is difficult to distinguish between them both. In summary, the results show that INDiC is able to determine the states of appliances well, so it is a possible approach for information extraction. In turn, INDiC can be used to evaluate the applicability of PACTS. We expect PACTS to significantly decrease the achieved disaggregation performance.

R.12 *PACTS has a significant influence on the INDiC disaggregation for the 'light' appliance, i.e., PACTS is able to keep INDiC from extracting information.* First, we inspect the result of applying noise for the discriminative pair s_{pair} = ('Light is in State 2','Light is in State 3'). After applying noise, the INDiC disaggregation results get worse (Fig. 3.8): Because s_{pair} should hide the distinction between States 2 and 3, we are particularly interested in the results covering the probabilities of both. An adversary having either State 2 or State 3 in mind obviously has difficulties in distinguishing which state is true: Guessing the right state is only 4% more likely than guessing the wrong one. The accuracy drops by 40% regarding State 2 and 23% regarding State 3, compared with the original INDiC results (Fig. 3.7). The results show that an adversary using INDiC can extract significantly less information regarding the 'light' appliance if PACTS is applied.

R.13 *PACTS has a significant influence on the INDiC disaggregation for State 2 of the 'refrigerator' and 'microwave' appliance, i.e., PACTS is able to keep INDiC from extracting information on that State. However, INDiC is still able to determine State 3 correctly in many situations.* Hiding the 'Refrigerator' State 2 and 3 leads to slightly different results: Determining State 2 is only correct in 6% of the cases. Guessing State 3 dropped by 36% but is still correct in 39% of the



(a) Confusion matrix

	State 1	State 2	State 3
State 1	*0.94*	0.04	0.02
State 2	0.16	*0.60*	0.24
State 3	0.48	0.06	*0.46*

(b) Predicted states *vs.* true states (relative)

Figure 3.7: INDiC results of 'Light'

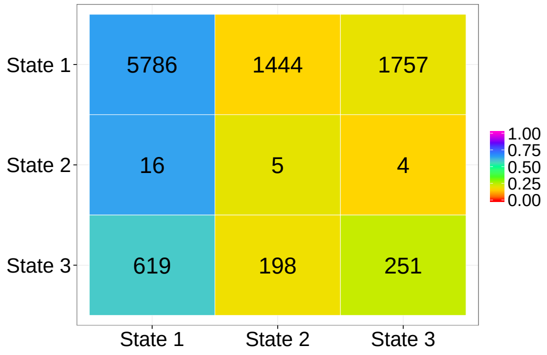
cases. Hiding the 'microwave' appliance leads to similar results. Although hiding State 2 works well, State 3 is still guessed correctly in approximately 52% of the cases (Fig. 3.12), with a drop of 41% compared with the data without noise (Fig. 3.11). In addition, guessing State 2 or 3 increases when State 1 was correct. This is an effect of the limitation discussed in Section 3.4.2 regarding consumption values lower than zero.

R.14 *PACTS is able to cover secrets induced by the INDiC approach.* Creating discriminative pairs for all appliances in Table 3.4 works similarly to the example in Section 3.4.2. Thus, PACTS is applicable when hiding information regarding the states of running appliances. Leaving the limitations regarding possible negative values aside, PACTS can cover all of the secrets induced by INDiC.

3.4.3 Generality: Re-identification

Re-identification, as introduced in Section 2.2, means linking personal data that do not contain any direct identifiers (e.g., name, address) to individuals. Reconsider that for the re-identification, an adversary has features regarding the individual power consumption as external knowledge. The distance (similarity) between the features of the time series in question and those in the external knowledge are computed. If the features are similar, i.e., the distance is 'low' the time series belongs most likely to the individual in the external knowledge.

Features of the consumption help to re-identify time series of power-consumption values. Changing such features in the published time series should prevent re-identification. We first analyze whether PACTS covers all of the features proposed by our re-identification method and then focus on the effects of hiding specific features. In particular, we focus on the following four features: over-

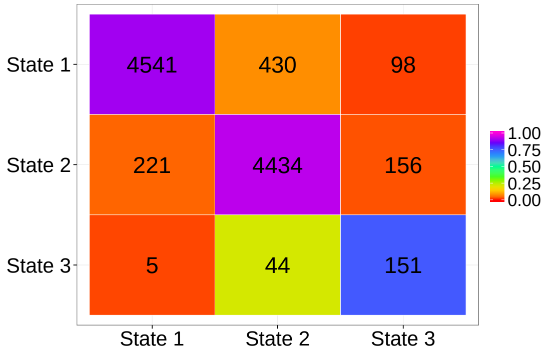


(a) Confusion matrix

	State 1	**State 2**	**State 3**
State 1	*0.64*	0.16	0.20
State 2	0.64	*0.20*	0.16
State 3	0.58	0.19	*0.23*

(b) Predicted states *vs.* true states (relative)

Figure 3.8: INDiC results of ‘Light’ with PACTS s_{pair} = (‘Light is in State 2’, ‘Light is in State 3’)

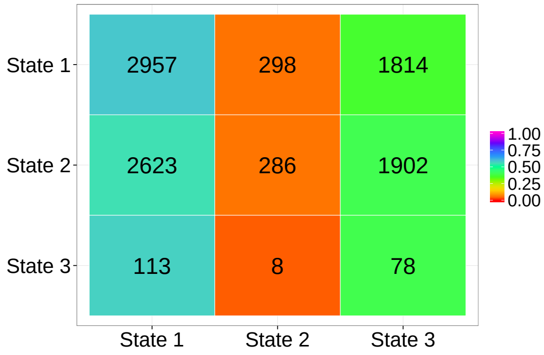


(a) Confusion matrix

	State 1	**State 2**	**State 3**
State 1	*0.90*	0.08	0.02
State 2	0.05	*0.92*	0.03
State 3	0.03	0.22	*0.75*

(b) Predicted states *vs.* true states (relative)

Figure 3.9: INDiC results of ‘Refrigerator’

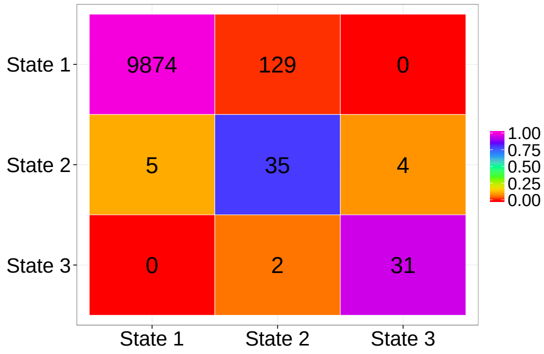


(a) Confusion matrix

	State 1	**State 2**	**State 3**
State 1	*0.58*	0.06	0.36
State 2	0.55	*0.06*	0.4
State 3	0.56	0.04	*0.39*

(b) Predicted states *vs.* true states (relative)

Figure 3.10: INDiC results of 'Refrigerator' with PACTS s_{pair} = ('Refrigerator is in State 2', 'Refrigerator is in State 3')

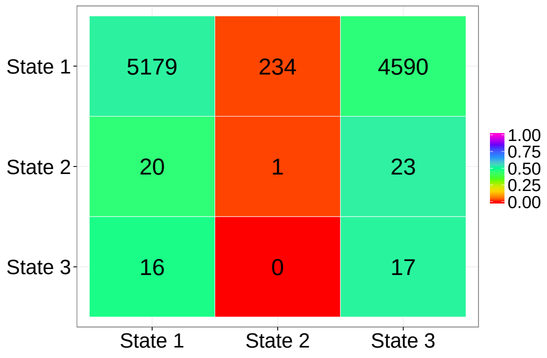


(a) Confusion matrix

	State 1	**State 2**	**State 3**
State 1	*0.99*	0.01	0.0
State 2	0.11	*0.79*	0.09
State 3	0.0	0.06	*0.93*

(b) Predicted states *vs.* true states (relative)

Figure 3.11: INDiC results of 'Microwave'



(a) Confusion matrix

	State 1	State 2	State 3
State 1	*0.52*	0.02	0.46
State 2	0.45	*0.02*	0.52
State 3	0.48	0.0	*0.52*

(b) Predicted states *vs.* true states (relative)

Figure 3.12: INDiC results of 'Microwave' with PACTS s_{pair} = ('Microwave is in State 2', 'Microwave is in State 3')

all consumption (Definition 5), maximum (Definition 7) and minimum consumption (Definition 6) for a time interval, and average bedtime hour (Definition 14).

Coverage of the Re-identification Features

Table 3.5 lists the necessary transformations and relevant coefficients for each feature of the re-identification. In this section, we explain in greater detail our decisions regarding the transformations.

Overall, Minimum, and Maximum Consumption: The sum and the minimum or maximum value of the position of the whole time series is in the y-dimension (consumption). The absolute height is reflected by the scaling coefficients of the wavelet-transformed representation. We explain the instantiation of secrets for these features in Section 3.4.3.

Standard Deviation and 0.9-Quantile: Both features depend on the spread of values in a time series. Applying noise to the Fourier-transformed representation changes the amplitudes of the contained frequencies. Obviously, the standard deviation and the 0.9-quantile also change.

Frequency of Mode: Applying noise to the contained frequencies in a signal also changes the frequency of each unique value.

Feature	**Transformation**	**Coefficients Concerned**
Overall Consumption (Def. 5)	Haar-Wavelet	Scaling Coefficient
Minimum Consumption (Def. 6)	Haar-Wavelet	Scaling Coefficient
Maximum Consumption (Def. 7)	Haar-Wavelet	Scaling Coefficient
Standard Deviation (Def. 8)	Fourier	All
0.9-Quantile (Def. 9)	Fourier	All
Frequency of mode (Def. 10)	Fourier	Significant Frequencies
Consumption Mo-Fr h_1-h_2 (Def. 11)	Decomposed Wavelet	Relevant Scaling Coefficient
Weekend Consumption (Def. 12)	Fourier	Frequencies Reflecting Fraction
Average Wakeup Hour (Def. 13)	Haar-Wavelet	Level 1/2
Average Bedtime Hour (Def. 14)	Haar-Wavelet	Level 1/2

Table 3.5: Feasible transformation for re-identification features

Consumption Mo-Fr $h_1 - h_2$: The decomposed wavelet transformation has to be applied in an aligned way to cover this feature. The part of the signal containing the consumption between h_1 and h_2 has to be transformed with the wavelet transformation in isolation. If the distance between h_1 and h_2 does not consist of 2^n values, this part also has to be decomposed. Applying noise to the scaling coefficients between h_1 and h_2 will change the consumption in the timespan in question analogue to the overall consumption.

Weekend Consumption: This feature considers the consumption during the weekend in relation to the consumption during the week. In the Fourier transformed representation this relation is covered by frequencies having the highest amplitude value during weekends, and their lowest amplitude value during weekdays. If noise is applied to those frequencies the relation between both consumptions change.

Average Wakeup and Bedtime Hour: The wakeup hour is the first significant increase in the morning and the bedtime hour the significant decrease at night. This is similar to the flow heater (see Example 5) and is reflected by the first and second level of the Haar wavelet–transformed representation.

In the following, we will elaborate on how the ‘Overall’, ‘Maximum’ and ‘Minimum Consumption’, and ‘Average Wakeup Hour’ and ‘Bedtime Hour’ features can be hidden with PACTS. We have chosen these features because they have the same structure as almost half of the features listed in Table 3.5.

Implementation of Secrets for Overall, Maximum, and Minimum Consumption

To implement secrets, we have to take a closer look at the feature definition in combination with PACTS. The overall power consumption of a time period is the sum of all channels $i \in [1, \ldots, n]$:

$$\sum_{\forall t \in \widehat{\mathcal{T}}} f[t] = \sum_{\forall t \in \widehat{\mathcal{T}}} f_1[t] + \cdots + \sum_{\forall t \in \widehat{\mathcal{T}}} f_n[t]$$

An adversary having external knowledge of the power consumption and trying to re-identify a record has to take inaccuracies into account, i.e., he or she typically does not know the total power consumption for sure, only within a certain range. Thus, we partition the channels into a known one, such as the relevant channel r, and the ones not known. The unknown channels are responsible for the difference between the known channels and the total consumption at each point of time.

$$\sum_{\forall t \in \widehat{\mathcal{T}}} f[t] = \sum_{\forall t \in \widehat{\mathcal{T}}} f_1[t] + \cdots + \sum_{\forall t \in \widehat{\mathcal{T}}} f_r[t] + \cdots + \sum_{\forall t \in \widehat{\mathcal{T}}} f_n[t]$$

Based on the sum $\sum_{\forall t \in \widehat{\mathcal{T}}} f[t]$, the adversary has to decide whether the known channel is consistent with his or her knowledge. Adding Laplace noise in line with ϵ-Pufferfish privacy leads to uncertainty regarding $\sum_{\forall t \in \widehat{\mathcal{T}}} f_r[t]$. Re-identification is successful if an adversary is able to single out the true individual record. In particular, this is relatively easy if the feature values of individuals are spread over a wide range and are rather unique. Thus, individual privacy requirements depend on assumptions regarding other individuals in the data set. Describing a suitable secret is deciding which interval is sufficient to hide $\sum_{\forall t \in \widehat{\mathcal{T}}} f_r[t]$ amongsother channels. We use the following notation:

$$s_k = \text{'Known power consumption is in interval [y-k, y+k]'}$$

The discriminative pairs can be of the form $s_{pair} = (s_k, s_{3k})$. One way to determine k is to look at the distribution of a known data set. Figure 2.2a indicates that $k = 5\ kWh$ is sufficient to hide a single household among more than 10 others for a large fraction of households. These considerations also hold for the 'Minimum' and 'Maximum' features.

Applying Noise to the Scaling Coefficient: Applying noise to the scaling coefficient is special, compared with other coefficients. In particular, the scaling coefficient is normed. It represents the overall, minimum, and maximum consumption, and is calculated as follows: $\frac{\sum_{\forall t \in \widehat{\mathcal{T}}} f[t]}{\sqrt{\|\widehat{\mathcal{T}}\|}}$. Thus, the additive noise $Laplace(4k/\epsilon)$ is also normalized: $\frac{\sum_{\forall t \in \widehat{\mathcal{T}}} f[t]}{\sqrt{\|\widehat{\mathcal{T}}\|}} + \frac{Laplace(4k/\epsilon)}{\sqrt{\|\widehat{\mathcal{T}}\|}}$.

Implementation of Secrets for Average Wakeup and Bedtime Hour

According to Definition 14 the bedtime hour is when a household switches off certain devices, e.g., the light or the television, right before going to bed. The devices do not necessary have to be the same for different households as long as they are usually switched off right before going to bed. We consider switch-off events only between 4 *p.m.* and 2 *a.m.* Some appliances may still run, but only the change in power consumption is of interest. An adversary trying to re-identify a household is interested in deciding whether the devices are switched off. Thus, an individual wants to hide the discriminative pair s_{pair}, which consists of the following secrets: s_1 = 'Household switches off devices before bedtime' and s_2 = 'Household does not switch off devices before bedtime'. The relevant channel r includes the devices mentioned for s_{pair}:

$$f_h[x] = f_h^r[x] + f_h^1[x] + \dots + f_h^n[x]$$

The switch-off causes a decrease in the power consumption of 0.5 kWh on $f_h^r[x]$. Thus, we apply Laplace$((4 \cdot 0.5)/\epsilon)$ noise on Level 1 and Laplace$((4 \cdot \frac{0.5}{\sqrt{2}})/\epsilon)$ noise on Level 2 during 4 *p.m.* and 2 *a.m.* Hiding wakeup hours is similar.

Results

The evaluation of PACTS as a way to hinder re-identification has two goals: First, we investigate whether PACTS is generally applicable, i.e., if we are able to formulate secrets regarding the features of re-identification. Second, we test how re-identification rates change if PACTS is applied, i.e., how effective PACTS is in hiding features from an adversary. The first result summarizes the general applicability.

R.15 *PACTS is able to hide all of the features used to re-identify households.* Section 3.4.3 explains how we are able to find an appropriate transformation for each of the proposed features. Thus, PACTS is able to express privacy requirements regarding re-identification.

To quantify effectiveness, we look at the relative decrease in accuracy, i.e., the number of households re-identified with and without applying noise. Although re-identification makes use of a combination of features to increase performance, to isolate the effects of hiding specific secrets we take only features relevant for the secret. In particular, we considered the following feature sets Φ grouped by different secret definitions: For secrets leading to noise on the scaling coefficient (Section 3.4.3) we choose $\{\phi^{MaxC}\}$, $\{\phi^{MinC}\}$, $\{\phi^{OC}\}$, and $\{\phi^{MaxC}, \phi^{MinC}, \phi^{OC}\}$. For secrets regarding the wakeup and bedtime hours, we use $\{\phi^{WH}\}$, $\{\phi^{BH}\}$ and the combination of both $\{\phi^{WH}, \phi^{BH}\}$. In total, we tested sets of 100, 500, and 1,000 persons living in 32, 158, and 314 households from the CER data set (see Section 2.2.3) and set $\epsilon = 0.1$. The parameters are similar to the ones chosen in the evaluation of the re-identification performance (see Section 2.2.4).

Choosing only a limited set of features reduces the number of re-identified households. This is the case both with and without applying noise, so this current evaluation is still conclusive. However, the effect of PACTS on the re-identification rate is difficult to quantify and is influenced by random effects when the rate is low on the unmodified data set. To isolate the effect of PACTS, we relax the re-identification condition in the following way: We deem a household as re-identified if the time series of the same household receives at least the nth lowest score. We choose $n = 1$

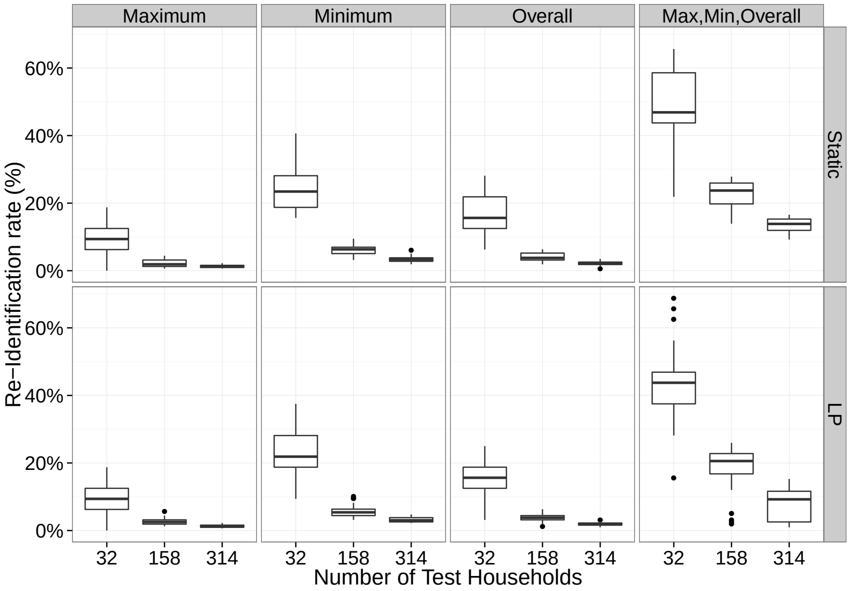


Figure 3.13: Re-identification rate for feature sets $\{\phi^{MaxC}\}$, $\{\phi^{MinC}\}$, $\{\phi^{OC}\}$, and $\{\phi^{MaxC}, \phi^{MinC}, \phi^{OC}\}$, $n = 1$

as the standard re-identification case and $n = 5$ as the relaxed case. Re-identification rates are significantly higher for $n = 5$ and for the combination of the features considered (Fig. 3.13–3.16).

Our measurements of the decrease in accuracy yields the following results:

R.16 *The accuracy decrease is more than 50% on average for features* $\{\phi^{MaxC}\}$, $\{\phi^{MinC}\}$, $\{\phi^{OC}\}$, *and their combination.* Applying PACTS significantly reduces the re-identification rate for $n = 1$ and $n = 5$ (Figs. 3.17 Figure 3.18), from at least 50% up to almost 100% on average. Thus, the proposed PACTS application is effective, as expected.

R.17 *The accuracy decrease is lower for feature sets consisting of* ϕ^{WH}, ϕ^{BH} *compared with feature sets consisting of* $\phi^{MaxC}, \phi^{MinC}, \phi^{OC}$. The lowest average re-identification rate drop for the wakeup and bedtime hour features approximately 25%, and the highest approximately 80%. This is lower than the decreases in Result **R.16** (see Figs. 3.19 and 3.20). The reason for the lower decrease is because the re-identification rate without applying PACTS was also lower. If the features of a number of households is similar, the algorithm starts to 'guess' the correct household. Consequently, a small number of households can still be re-identified by random guessing, leading to a lower decrease if the reference rate was also low.

We expected a decrease in re-identification performance as presented in Results **R.16** and **R.17**. However, the data also show some anomalies, which we will discuss in the following.

R.18 *Re-identification rate increases when applying PACTS in rare cases; however, this is not*

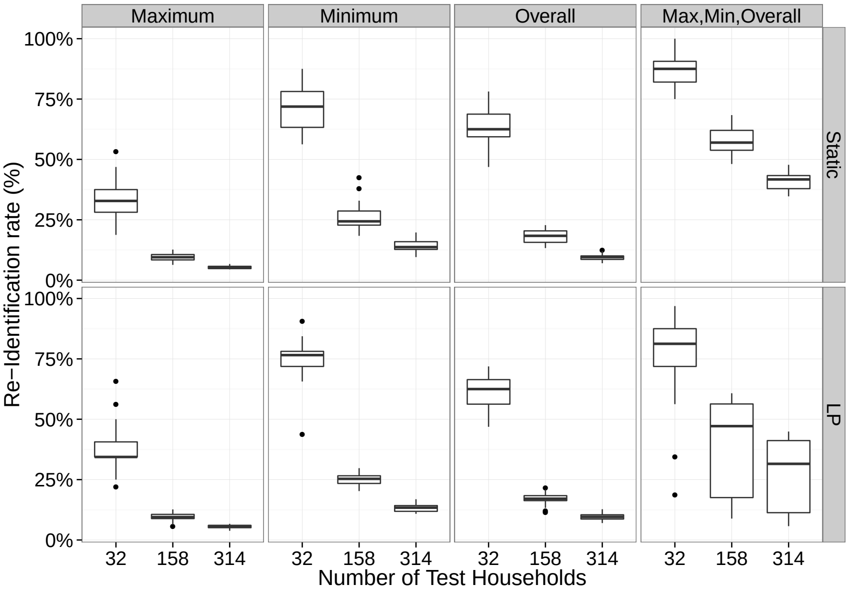


Figure 3.14: Re-identification rate for feature sets $\left\{\phi^{MaxC}\right\}$, $\left\{\phi^{MinC}\right\}$, $\left\{\phi^{OC}\right\}$ and $\left\{\phi^{MaxC}, \phi^{MinC}, \phi^{OC}\right\}$, $n = 5$

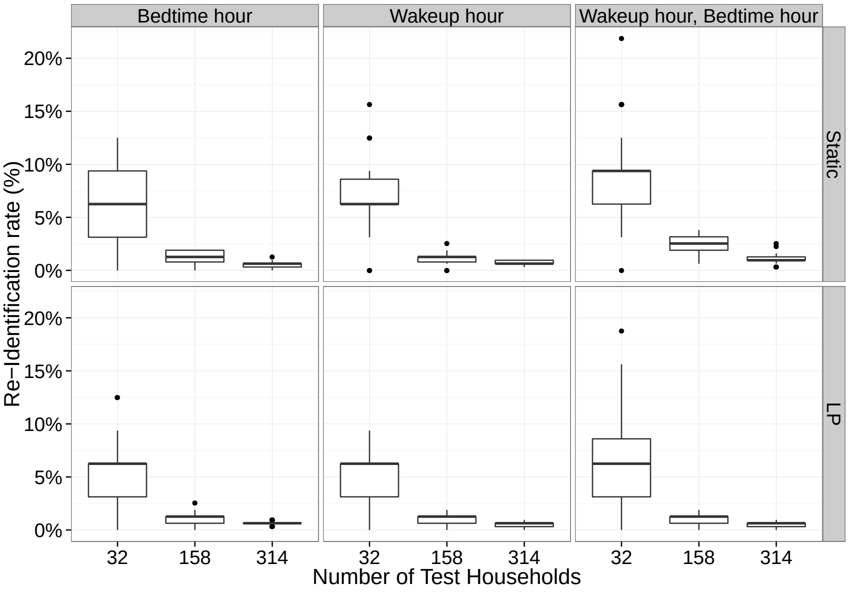


Figure 3.15: Re-identification rate for feature sets $\left\{\phi^{WH}\right\}$, $\left\{\phi^{BH}\right\}$ and $\left\{\phi^{WH}, \phi^{BH}\right\}$, $n = 1$

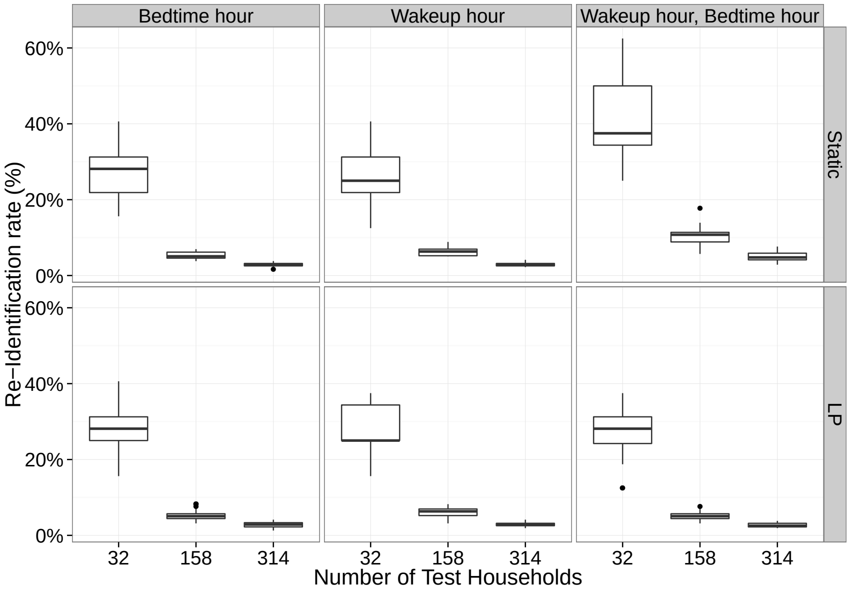


Figure 3.16: Re-identification rate for feature sets $\left\{\phi^{WH}\right\}$, $\left\{\phi^{BH}\right\}$ and $\left\{\phi^{WH}, \phi^{BH}\right\}$, $n = 5$

a limitation on the effectiveness. Reconsider the discussion of Result **R.17**: Random guesses are still possible and do not influence the general effectiveness of the proposed privacy-enhancing method. Additionally, in our evaluation, we have assumed the same discriminative pair for all households. However, for outliers in particular, e.g., a household consuming a lot of electricity and thus being easy to re-identify, discriminative pairs should differ. In particular, the k of the interval must be larger. A larger n and a lower re-identification rate are beneficial for the importance of (correct) random guesses and the influence of outliers to our results.

In summary, PACTS allows the definition of suitable secrets to hinder re-identification. Even with secrets designed in a straightforward way, without considering outliers, the accuracy decreases significantly.

3.5 Conclusions

Disclosure of data plays a significant role in the context of the smart grid. However, time series of smart-meter data contain sensitive information, which is represented in many different ways. Individuals may not allow access to the data as long as sensitive information based on individual privacy preferences is not removed, i.e., they require respect for their right to informational self-determination. Pufferfish is a state-of-the-art approach to hide specific information. However, application-specific work is required when applying it to smart-meter data and carrying out an

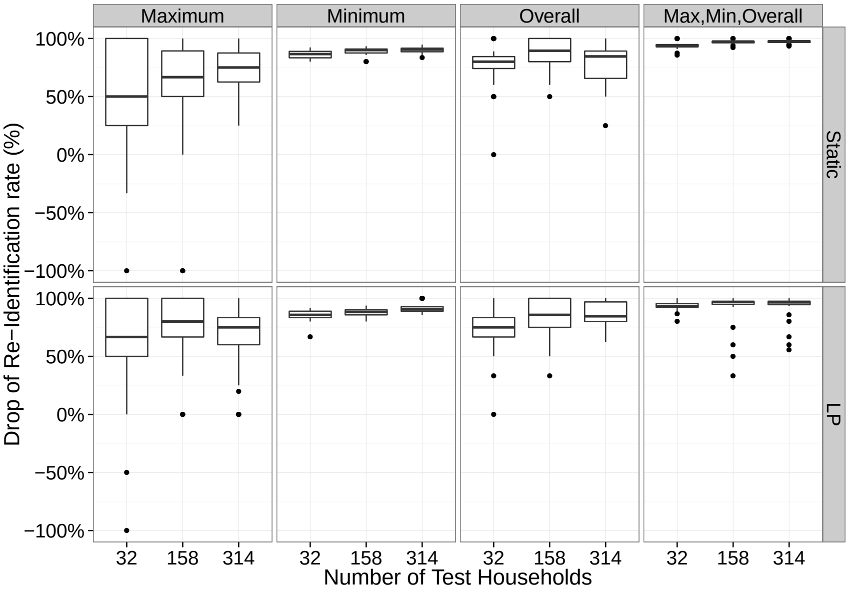


Figure 3.17: Drop in re-identification rate for feature sets $\{\phi^{MaxC}\}$, $\{\phi^{MinC}\}$, $\{\phi^{OC}\}$ and $\{\phi^{MaxC}, \phi^{MinC}, \phi^{OC}\}$, $n = 1$

evaluation that is conclusive. PACTS as a provable privacy approach for time series that include the definition of how sensitive information is represented, how data-evolution scenarios can be applied, and how the information can be perturbed to give Pufferfish guarantees. Next, it is challenging to evaluate the general coverage of secrets. Our evaluation has addressed this point: The evaluation has shown that PACTS can indeed shield personal information from information extraction or re-identification approaches. The potential of an adversary to gain information from the disclosed data set has dropped significantly.

PACTS can be applied to time series in general if the following holds: First, for the secrets in question, an abstracted representation with meaningful coefficients can be found. Within the abstracted representation, application of noise guarantees Pufferfish privacy, e.g., because smart-meter data are always an aggregate of several appliances, and additive noise is applicable. Similar structures can be found in other time series, e.g., time series of GPS trajectories. Reaching a certain location is the sum of several 'sub-movement' trajectories. For example, going to work involves a number of sub-movements because the individual uses several means of transportation.

With the help of privacy-enhancing methods such as PACTS, individual privacy preferences can be respected. However, all approaches have in common that the sensitive data are modified. This puts utility of the data at risk. To investigate whether the data are still useable, we require a measure for privacy-enhancing methods. We introduce such a measure in the next chapter.

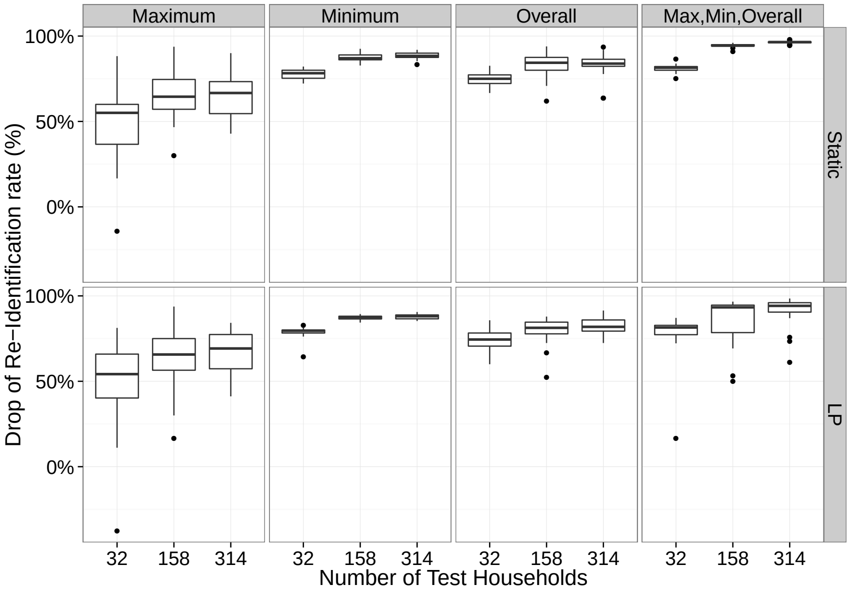


Figure 3.18: Drop in re-identification rate for feature sets $\left\{\phi^{MaxC}\right\}$, $\left\{\phi^{MinC}\right\}$, $\left\{\phi^{OC}\right\}$ and $\left\{\phi^{MaxC}, \phi^{MinC}, \phi^{OC}\right\}$, $n = 5$

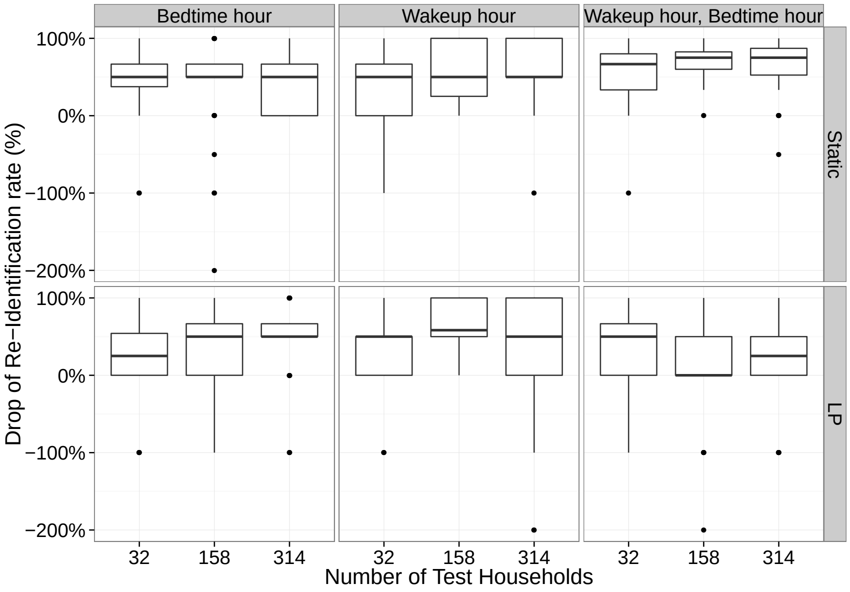


Figure 3.19: Drop in re-identification rate for feature sets $\left\{\phi^{WH}\right\}$, $\left\{\phi^{BH}\right\}$ and $\left\{\phi^{WH}, \phi^{BH}\right\}$, $n = 1$

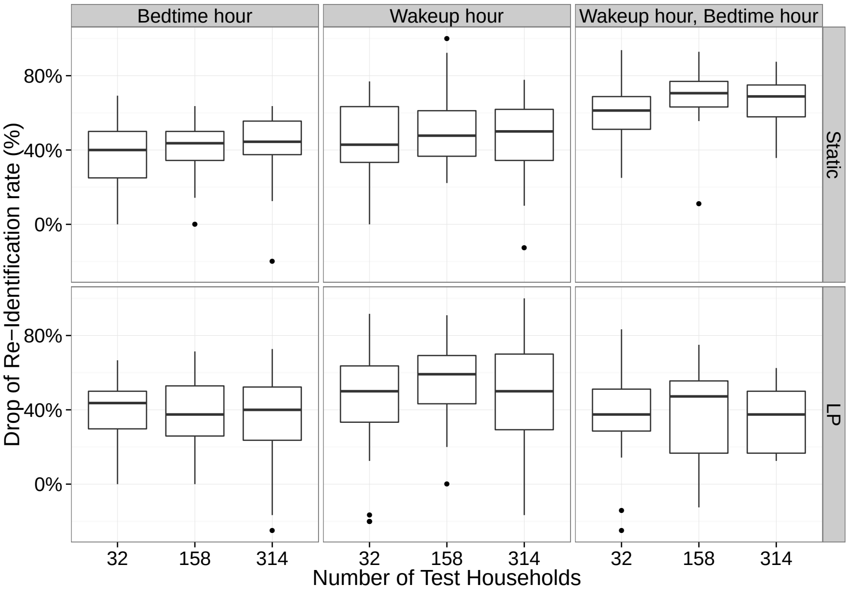


Figure 3.20: Drop in re-identification rate for feature sets $\{\phi^{WH}\}$, $\{\phi^{BH}\}$ and $\{\phi^{WH}, \phi^{BH}\}$, $n = 5$

Chapter 4

Local Energy Market: Application-Specific Data Quality Measure

Privacy-enhancing methods (see Chapter 3) have in common that they modify the actual time-series data. All methods strive to remove personal information to publish data without setting the privacy of individuals at risk. For example, a privacy-enhancing method for smart-meter time series modifies the power-consumption values. Methods differ in modifications of the data set, and stricter privacy guarantees require greater modifications of the data.

However, those modifications have a negative impact on the quality of the published data. This affects legitimate applications using such data sources. These applications provide results for the benefit of society. For instance, applications in the smart grid domain may improve reliability of supply or reduce emissions. From an application perspective, the ideal case is accessing unmodified data to provide the highest utility and thus the most benefit for society. Because the individual privacy interests compete with the common and legitimate interests of society, it is worthwhile investigating the tradeoff between privacy and utility [70].

Utility measures quantify the impact on data quality. We distinguish between *abstract* and *application-specific* utility measures. Abstract measures generally quantify the 'change' of the privacy-enhanced data in comparison with the original data. Popular examples are distance measures like the Euclidean distance. However, such measures do not necessarily quantify the actual effect on an application. We illustrate this in the following example.

Example 9 ***(Abstract Utility Measure)*****:** Suppose that a time series is perturbed two times. Furthermore, with the second perturbation, the Euclidean distance of the resulting series to the original one is twice as large as the first one. This does not necessarily mean that the utility is halved. For example, it may still be possible to identify outliers in the time series. □

Application-specific measures, in turn, quantify the effects of data modifications on concrete applications. The impact is quantified by the different results that an application yields on privacy-enhanced data in comparison with the unmodified data. Such measures allow a prediction of the influence that a privacy-enhancing method has on a specific application.

Example 10 ***(Application-Specific Utility Measure)*****:** Assume that the application in question is calculating the average of a time series. A specific measure for this application is the absolute difference between the average of a privacy-enhanced time series and the true average. In turn, this result does not allow inference on the Euclidean distance between the privacy-enhanced and the original time series. □

Challenges for a data-quality measure for privacy-enhancement methods are as follows:

Comprehensible: The PACTS (Section 3.2) privacy approach assumes that the individuals specify their privacy requirements. Furthermore, if their privacy is respected, individuals are willing to provide access to their data. Studies like [47] have shown that monetary incentives have a positive effect on disclosure. Thus, privacy requirements may change if individuals can understand the impact of privacy enhancement on utility and have an incentive to provide more accurate data. Consequently, our first challenge is that the resulting impact measures are comprehensive, i.e., the results of an impact measure for privacy-enhancement methods are intuitively understandable.

Discriminative for Applications: The goal of impact measurements is to provide insight on how well applications of the specific domain work with privacy-enhanced data. This is also important for an individual when an application accessing the privacy-enhanced data leads to a benefit. Thus, our second challenge is to express this correlation: The results of a utility measurement are discriminative for an application, i.e., they quantify the performance decrease and imply an order for different privacy-enhancement methods and parameters.

We propose an application-specific utility measure in the smart grid scenario based on an electronic local energy market. The proposed measure is both comprehensible and discriminative for applications. In particular, we investigate the impact of privacy-enhancing methods on an (electronic) local energy market (**C.3**). In a nutshell, it is an electronic marketplace for energy that is de-centrally produced. Privacy-enhancement methods that modify the bids of consumers lead to less efficient allocations. With the proposed market we can explore the economic effects of different privacy-enhancement methods and their parameters. As such, it offers a concrete application scenario to evaluate the actual impact of recent privacy-enhancement methods. Economic effects can be quantified in a comprehensive manner and units, making the proposed measure understandable.

To underline the validity of the proposed utility measure, we also contribute an in-depth analysis of privacy-enhancement methods that are applied in local energy markets in Chapter 5. The remainder of this chapter is structured as follows. We begin by discussing related impact measures (Section 4.1) and continue with the description of the local energy market as an application-specific utility measure (Section 4.2). We introduce our proposed model of such a market in Section 4.3

and describe our instantiation in Section 4.4. Section 4.5 concludes the chapter. The contents of this chapter are published in [12] and [57].

4.1 Related Impact Measures

In this section, we give a brief review of existing related impact measures for privacy-enhancing methods. We describe statistical measures, information theoretic approaches, and application-specific measures.

The standard deviation is a measure for uncertainty, i.e., a high standard deviation indicates that the considered values have a large difference to the reference point. [88] uses a standard deviation that considers the modified time-series value $f'(t)$ in comparison to the original value $f(t)$ as a reference point for each time slot. Differential privacy approaches for time series of smart-meter data also use the standard deviation (or related measures like the Root Mean Square error) of the error quantifying utility [5, 99]. This measure is only discriminative for applications if the result is entirely based on the deviation.

Privacy-protecting approaches using a rechargeable battery [54, 96, 112] use mutual information to quantify privacy and utility. Mutual information measures the information theoretic dependence between two variables. If there is a correlation between the original and the privacy-enhanced time series, the mutual information between both series is large. This indicates that the utility is also large. However, to understand this measure, a information theoretic background is necessary. Thus, it is not intuitively comprehensible.

Approaches for the time series of trajectory data use distance measures of time series such as the log cost metric and the Euclidean distance [86, 94]. As shown in Example 9, such measures do not necessarily discriminate applications. Regarding trajectory data, one common information need are range queries that result in objects nearby a specific query point. [1, 2, 120] quantify utility with range query distortions. [20] follows a similar approach, but with different queries regarding frequent sequences and doublets. Because such queries are commonly used, they are comprehensible and discriminate applications. However, they cover only a specific fact in isolation. The proposed measures in the local energy market (Section 4.2) are influenced by a combination of different ‘queries’ covering the supply and the demand side. This allows more general conclusions on the utility of smart-meter data. To our knowledge, an application-specific-utility measure in the smart grid with similar complexity does not exist.

4.2 Application Specific Impact Measure: An Electronic Local Energy Market

Historically, the electricity grid is tailored to a centralized generation structure. At its core are a few large power plants that generate electricity for a large number of consumers [37]. However, reducing the CO_2 emissions of the energy production requires the integration of renewable sources, such as photovoltaic sites and micro-combined heat and power plants. These sources are volatile and distributed. Compared with a large power plant, each of them produces only a small amount of power and cannot be controlled centrally. Because of their variability, integration of renewables remains a big challenge for today's power system. See Section 1.1 for more details.

Smart grids [77], the ICT-enabled electricity networks of the future, facilitate new operational paradigms [41, 97]. A case in point is the establishment of local energy markets, which provide a way to match regional energy demand and renewable supply [48, 65]. More 'local' (i.e., in spatial proximity of generation) energy consumption can help to improve integration of renewables and minimize transmission losses [4]. To work efficiently, local energy markets rely on truthful power-consumption information revealed by the participants, e.g., private households. In such markets, customers cover their energy needs by bidding for the required energy amounts over short time intervals. Consequently, a customer's consumption behavior is encapsulated in these bids.

We have seen in Chapter 2 that fine-grain power consumption data contain a vast variety of different personal information. Consequently, electronic market systems in smart grids should strive to preserve privacy properties [78]. We have introduced privacy-enhancement methods in Chapter 3. Their common ground is the distortion of sensitive values, i.e., energy consumption levels. In doing so, they are able to retain a certain level of privacy despite personal data being revealed to the market.

However, distorted bids are likely to induce less efficient allocations. Depending on the nature of the distortion, more or less energy than actually needed may be allocated. Hence, privacy enhancement may lead to additional costs for consumers. As an utility measure, we quantify these privacy costs in a local energy market with demand-side flexibility and storage.

To understand the relationship of privacy enhancement and local energy markets, we model a smart grid marketplace with privacy-enhancement methods together with a customer demand model. We characterize customer-bidding behavior and determine formal characteristics of the interplay between components of our model. In particular, we provide the following:

- The personal data flows for a local electricity market based on a double-sided auction;
- A model for the supply and demand of electricity in a low-voltage circuit in the near future based on realistic data;
- An investigation on how privacy-enhancing approaches can be applied to order books and ensure the privacy of participants; and
- A characterization of customer-bidding behavior under different assumptions. We also include different pricing schemes, that reflect traditional and 'smart grid'-enabled utility of consumers.

4.2.1 Related Work Considering Local Energy Markets

In this section, we review related local energy market approaches, market transparency issues including cryptographic approaches, and encompass our approach for using storage systems with existing privacy protection approaches.

Local Energy Markets Renewable sources for electricity generation are distributed and volatile by nature. The efficient use of such sources is an important part of the smart grid vision. Local energy markets efficiently coordinate the decentralized generation of electricity [48]. Generators of renewable energy and consumers participate in such local markets and trade energy over short time intervals, e.g., 30 min or less. Local markets for renewable energies have a number of positive effects: Community-based funding models increase the acceptance and accelerate the installation of renewable power plants [52, 122]; local markets seamlessly fit into the distributed structure of renewable power generation [48]; and finally, because the price of renewable energy is falling below that of conventional energy, private and shared consumption are becoming more important [101].

Transparency and Disclosure Obligations Transparency obligations like the EUC 543/2013 mandate the publication of comprehensive market data. Market transparency is key to ensure market liquidity and hence market efficiency [76, 87]. To cope with the over- and undersupply of renewable energies, distribution system operators and plant operators must be able to forecast the energy demand and the production of renewable sources [113]. Thus, the production and allocation of renewable energy sources should be openly known.

Cryptographic Auctions Cryptographic auctions encrypt the bids and provide verifiability of the correctness. However, they do not allow ex-post access to the information, which limits market transparency. Furthermore, they do not facilitate repeated and parallel market interactions as they are designed for a single seller [10], or preserve secrets only until the end of an auction [89].

Storage and Privacy Privacy protection approaches like [54, 96, 112] rely on the use of energy storage. See Section 3.1.2 for more details. In both cases, a stationary storage is used to completely mask the load signatures of the underlying household appliances. However, these results are primarily anecdotal and rely on an arbitrarily large storage system. In the proposed local energy market we follow the general idea by investigating the economic interplay between privacy enhancement and a fixed storage system with limited capacity. This allows us to compare the previously orthogonal dimensions of storage costs and privacy.

4.3 Model

In this section, we specify our theoretical local energy market model and the corresponding privacy-enhancement methods. This includes details on the bidding process and definitions of the market-performance measures.

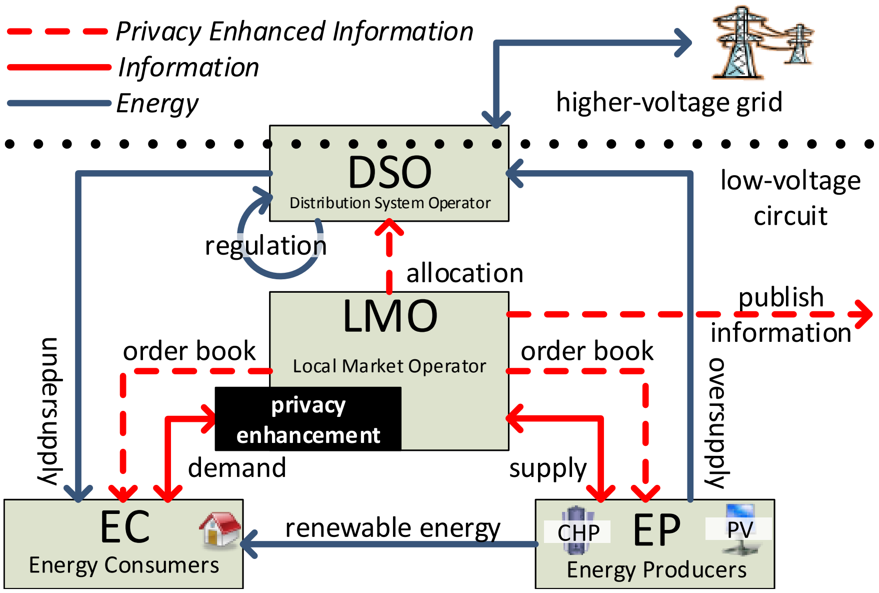


Figure 4.1: Architecture of our privacy-aware electronic local energy marketplace

4.3.1 Technical Architecture

Before discussing the proposed market model, we clarify how such a market can be realized. Local renewable energy sources such as photovoltaic roof installations are connected to the power grid on the level of low-voltage (below 1500 Volts) distribution circuits. Typically, a low-voltage circuit distributes electricity in an urban district or in a village [50]. Because a step-up transformation to higher voltages for long-distance lines would decrease energy efficiency, the local low-voltage circuit specifies the participants for our local energy market [49]:

- **Distribution System Operator (DSO)** In our market, the DSO balances over- and undersupply of locally renewable energy. For example, if the photovoltaic installations cannot meet the demand at night, the DSO transfers energy from the higher voltage grid levels. In the case of oversupply, the DSO exports surplus energy.

- **Energy Producer (EP)** Each EP offers an amount of energy for a future period of time. Because larger sources, e.g., wind turbines, are usually connected to higher-voltage grids, we consider only micro-combined heat and power plants (CHP) [25] and small photovoltaic sites (PV) [91].

- **Energy Consumer (EC)** The EC places bids for energy at certain time intervals. The typical ECs on a low-voltage circuit are private households. This is realistic, as large enterprises are connected to the higher-voltage grid.

- **Local Market Operator (LMO)** The LMO matches the demand of the ECs with the supply from the EPs. To do so, the LMO manages an order book containing the bids from the ECs and EPs. The order book is public for all market participants and the DSO. To ensure privacy, the LMO applies privacy-enhancement methods to the bids of the consumers.

Figure 4.1 depicts our architecture. Each EC predicts its future energy demand. Likewise, the EPs predict their energy production, e.g., based on the weather forecast. The ECs and EPs then place bids for a certain amount of energy at a certain period of time. The LMO computes the outcome of the auction and communicates this information to the EC, EP, and DSO. Each EP feeds all energy from its plant to the local low-voltage circuit.

Considering a real-world scenario, situations with over- and undersupply have to be handled. The DSO is capable of importing and exporting electricity from higher voltage grids, guaranteeing security of supply. Each EC is charged for electricity that is provided locally and for conventional energy drawn from the DSO. In summary, local energy markets are applicable in a real-world scenario. In the following we specify the market model from an economical perspective.

4.3.2 Market Structure

We want to study the effects of applying privacy-enhancement techniques on such a local energy market with the help of a model. Following the market engineering paradigm [65, 116], an appropriate model of a marketplace requires specifying the participants and their behavior (agent behavior), the transaction object, the market mechanism (market microstructure), and market performance measures (market outcome). We further describe the integration of privacy-enhancement methods in such a structure.

Market Participants **Notation 10 (Market Participants):** $\mathcal{A}$ is the set of participants in our local energy market. $\mathcal{C}$ denotes the set of consumers, and $\mathcal{G}$ the set of producers.

Each participant (actor) $a \in \mathcal{A}$ is either a consumer $c \in \mathcal{C}$ or a producer (generator) $g \in \mathcal{G}$, i.e., we assume $\mathcal{C} \cap \mathcal{G} = \varnothing$. We do not consider prosumers (producer and consumer) as they give rise to new strategic considerations by acting on both sides of the market.

Consumer energy demand varies over time. We model the time domain as a set of time intervals $\mathcal{T}$(see Notation 1). For each $t \in \mathcal{T}$, a consumer's maximum consumption level, referred to as the saturation level, is given by $\overline{x}_c(t)$. The trajectories of saturation levels form a set of time series: $\mathcal{X} = \{\overline{x}_c(t) | t \in \mathcal{T}, c \in \mathcal{C}\}$. The purchasing behavior of a consumer is governed by individual utility as specified in Section 4.3.3. Given temporally varying electricity needs $\overline{x}_c(t)$, optimal bidding requires customers to dynamically determine quantity-utility mappings. The notation in the market context follows the common notation in Section 2.2.1.

The set of producers consists of local generation units (PV, CHP) and a balancing party. This party is run by the DSO and reflects energy imports from the superordinate grid. Producers participate in the market by selling electricity. Individual rationality requires them to at least cover their marginal generation costs. Their capacity limits their bid quantities.

Transaction Object Our market instantiation follows traditional wholesale electricity markets in that electrical energy supply and demand commitments are traded. A bid contains the issuing market participant and its type (buy or sell order), the amount of energy procured (in kWh), and the reservation price p_{lim}. In the case of a sell order, the latter specifies the minimum price; in the case of a buy order, it is the maximum price. Individual actors can submit several bids to reflect nonlinear customer utility and generator cost functions.

Notation 11 (Order): An order o_a of market participant a is a triple containing the quantity (q), the limit price (p_{lim}), and the actual allocated amount (q_{alloc}): $o_a = (q, p_{lim}, q_{alloc})$. If we specifically address an element of this triple, we refer to it as $o_a[q]$,$o_a[p_{lim}]$ and $o_a[q_{alloc}]$. Because we omit prosumers, o_c is necessarily a buy order and o_g a sell order.

Privacy-Enhanced Bidding As noted previously, a customer c will formulate his or her bid at time t to reflect his or her current energy-demand saturation level $\overline{x}_c(t)$. Consumers will place a collection of bids reflecting their utility function under $\overline{x}_c(t)$. In the presence of privacy enhancement, the bidding process slightly changes: Consumers report $\overline{x}_c(t)$ to the privacy protection system, which in turn determines a modified demand report $\widetilde{x}_c(t)$ that is communicated to the market.

There are two distinct elements in the report that a consumer could strategize about: quantity and price. As private information is embedded only in the quantity component, we rule out that consumers will modify their demand report as this could open a side channel that undermines the privacy-enhancement technique. Consequently, we assume that the privacy protection system receives the true initial demand reports of the consumers. For the valuation, we do not make this assumption but show that in specific cases privacy-aware auctions are incentive compatible with respect to prices and consumers, and thus will optimally report their true valuation to the system. See Lemma 10 for further discussion. Bidding with and without a privacy-enhancement method is illustrated in Figure 4.2.

Market Mechanism Because electricity is a homogeneous good, double-sided auction formats can achieve a high level of market liquidity and efficiency [49, 98]. Following previous research on local energy markets, e.g., [65], we select the discrete time double auction (also referred to as periodic call auction or call market) as the market mechanism. Here, market clearing is not continuous but occurs in repeated time slots $t \in \mathcal{T}$. For each time slot, the market mechanism determines the allocation and clearing price for the submitted bids and asks. It does so by first constructing demand and supply curves and subsequently determining the intersection of the two. All of the necessary orders are contained in the order book. Formally, we denote the order book the following way.

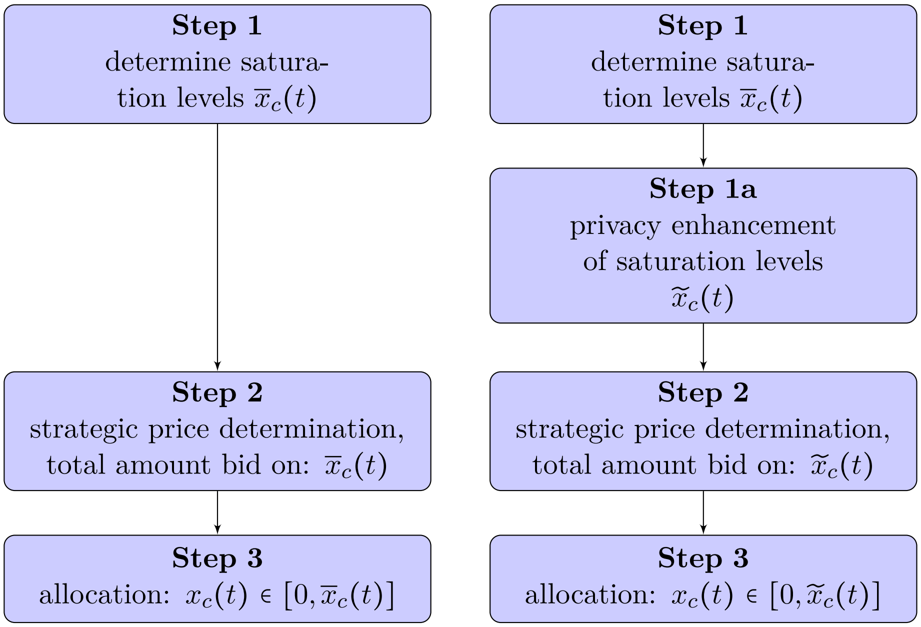

Step 1
determine saturation levels $\overline{x}_c(t)$
Step 2
strategic price determination, total amount bid on: $\overline{x}_c(t)$
Step 3
allocation: $x_c(t) \in [0, \overline{x}_c(t)]$
Step 1
determine saturation levels $\overline{x}_c(t)$
Step 1a
privacy enhancement of saturation levels $\widetilde{x}_c(t)$
Step 2
strategic price determination, total amount bid on: $\widetilde{x}_c(t)$
Step 3
allocation: $x_c(t) \in [0, \widetilde{x}_c(t)]$

Figure 4.2: Auctions *without (left)* and *with (right)* privacy enhancement

Notation 12 (Order Book): The order book $\mathcal{O}_t$ contains all orders for time slot t. We refer to the sell orders as $\mathcal{O}_t^s$ and to the buy orders as $\mathcal{O}_t^b$. Furthermore, the order book $\widetilde{\mathcal{O}}$ contains the same orders, but the bids of the consumers are replaced by their privacy-enhanced bids.

Orders match if the limit price of the sell order is lower than the one of the buy order, and if the quantity is not completely allocated. The double auction mechanism determines the clearing price p_t for each time slot by the limit prices of the last matched orders. We use a uniform pricing scheme (k-pricing) in which the clearing price is between the limit prices of the last matched orders. We set $k = \frac{1}{2}$ and the clearing price is the average of the limits. Algorithm 4 contains a pseudo-code implementation of the double-sided auction, and Figure 4.3 illustrates the sample order book in Table 4.1. For a more detailed description of the call market, we refer to [90].

Market-Quality Measure To assess the economic outcome of our local marketplace, we analyze the market's allocative efficiency. Consequently, we use social welfare as an application-specific measure for the effect of privacy-enhancement mechanisms.

Definition 32 (Social Welfare): Social welfare is the sum of consumer surplus (difference between willingness to pay and clearing price) and producer surplus (difference between clearing price and costs):

Algorithm 4: Discrete Time Double Auction with k-pricing ($k = \frac{1}{2}$)

```
Input: Order book for time slot $t$: $\mathcal{O}_t = \mathcal{O}_t^b \cup \mathcal{O}_t^s$
Result: Allocation for time slot $t$ and clearing price $p_t$
List sellOrders = sortAscendingByLimit($\mathcal{O}_t^s$);
List buyOrders = sortDescendingByLimit($\mathcal{O}_t^b$);
Order $o_g$ = sellOrders.first();
Order $o_c$ = buyOrders.first();
// stop if orders do not match
while $o_g[p_{lim}] \leq o_c[p_{lim}]$ do
    // match the remaining order volumes
    volume = min($o_g[q]$ - $o_g[q_{alloc}]$, $o_c[q]$ - $o_c[q_{alloc}]$);
    $o_g[q_{alloc}] = o_g[q_{alloc}]$ + volume;
    $o_c[q_{alloc}] = o_c[q_{alloc}]$ + volume;
    // fetch next unmatched orders if necessary and possible
    if $o_g[q] == o_g[q_{alloc}]$ then
        if not sellOrders.hasNext() then break;
        $o_g$ = sellOrders.next();
    end
    if $o_c[q] == o_c[q_{alloc}]$ then
        if not buyOrders.hasNext() then break;
        $o_c$ = buyOrders.next();
    end
end
// compute price from last matching orders
$p_t = \frac{o_g[p_{lim}] + o_c[p_{lim}]}{2}$;
```

Trader	Sell	Limit	Buy
PV2	3 kWh	0.250 €/kWh	
PV1	5 kWh	0.210 €/kWh	
CHP2	3 kWh	0.190 €/kWh	
PV3	4 kWh	0.160 €/kWh	
HH1		0.240 €/kWh	3 kWh
HH2		0.235 €/kWh	5 kWh
HH3		0.230 €/kWh	1 kWh
HH4		0.220 €/kWh	6 kWh

Table 4.1: Order book for single time slot

$$\mathcal{W} = \sum_{\forall c \in \mathcal{C}} CS_c + \sum_{\forall g \in \mathcal{G}} PS_g$$

□

To ease the comparison of different simulation runs, we rely on the relative welfare.

Definition 33 (Relative Welfare): Let $\mathcal{W}$ be the welfare achieved in a local energy market without privacy enhancement. Furthermore, let $\widetilde{\mathcal{W}}$ be the welfare achieved in the same market (concerning supply and demand), but in the presence of a privacy-enhancement method that modifies the bids of consumers. Relative welfare $\mathcal{W}'$ is then given by

$$\mathcal{W}' = \frac{\mathcal{W}}{\widetilde{\mathcal{W}}}$$

□

We posit that higher relative welfare is an indication that a privacy-enhancement method that retains a higher data quality in the application scenario under consideration.

Local energy markets strive to reduce CO_2 emissions. Locally traded energy is renewable and does not produce emissions. If consumers demand energy from the higher grid level, emissions depend on the mix of primary sources at a specific time of the day. The definition of CO_2 intensity enables us to quantify the saved emissions.

Definition 34 (CO_2 intensity): The CO_2 intensity is the amount of CO_2 emitted per consumed kWh for a specific time slot t. We denote that as $\omega(t)$. Intensity is measured in $\frac{gCO_2}{kWh}$.
□

Intuitively, the saved emissions is the demand fulfilled locally instead of by requesting electricity from higher grid levels, respectively the balancing party.

Definition 35 (Saved Emissions): Assume that the order book $\mathcal{O}$ contains all orders and allocated amounts. Let $D(c, t, \mathcal{O}_t)$ be the total fulfilled demand of consumer c at time slot t, defined

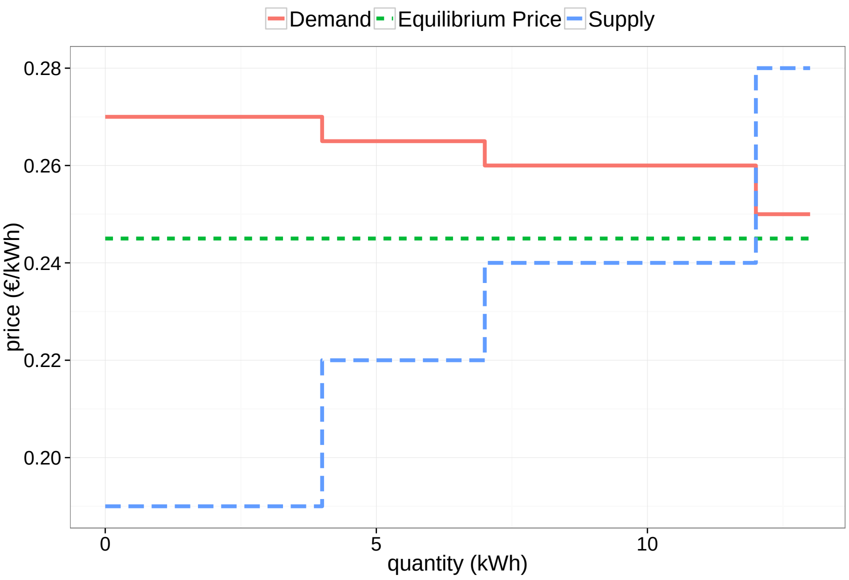


Figure 4.3: Illustration of a double auction

as follows:

$$D(c,t,\mathcal{O}_t) = min(\overline{x}_c(t), \sum_{\forall o_c \in \mathcal{O}_t} o_c[q_{alloc}])$$

The total sum of the allocated electricity can exceed the saturation level, as privacy enhancement may lead to a higher saturation level $\widetilde{x}_c(t)$. Then the saved emissions $\Omega(\mathcal{O})$ are defined as follows:

$$\Omega(\mathcal{O}) = \sum_{\forall t \in \mathcal{T}, \forall c \in \mathcal{C}} D(c,t,\mathcal{O}_t) \cdot \omega(t) - \sum_{\forall t \in \mathcal{T}} \{o_b[q_{alloc}] \cdot \omega(t) | o_b \in \mathcal{O}_t \wedge b \in \mathcal{G} \text{ is balancing party}\}$$

□

As a market-quality measure, we quantify the relative saved emissions. Because privacy-enhancement allocative inefficiency leads to additional emissions, the higher the relative saved emissions, the more successful the market.

Definition 36 (Relative Saved Emissions): Let order book $\mathcal{O}$ contain all orders and allocated amounts. Furthermore, let $\widetilde{\mathcal{O}}$ be the corresponding order book based on the privacy-enhanced bids. The additional emissions Ω' are defined as follows:

$$\Omega'(\mathcal{O}, \widetilde{\mathcal{O}}) = \frac{\Omega(\widetilde{\mathcal{O}})}{\Omega(\mathcal{O})}$$

□

4.3.3 Customer Model

The key element in modeling customer interactions (i.e., bidding behavior) with the market is the underlying utility model. Although electricity traditionally is subject to billing and is considered a homogeneous good, the smart grid includes differentiated energy services [102], and we follow this notion. To this end, we propose an analytical customer model similar to [9].

Customer Utility A costumer c can place a number of orders for a timeslot. Depending on the supply and limit prices on the market, not all orders may be completely fulfilled. The total allocated amount for a consumer is denoted as follows.

Notation 13 (Consumer's Allocated Electricity for Time Slot t): For time slot t a customer has allocated $x_c(t)$ electricity in total. This is the sum of allocations of the orders:

$$x_c(t) = \sum_{\forall o_c \in \mathcal{O}_t} o_c[q_{alloc}]$$

The definition of the customer's utility completes the model.

Definition 37 (Utility): The customer utility is a nondecreasing concave function, defined for each customer $c \in \mathcal{C}$ and for each time slot $t \in \mathcal{T}$.

$$U_{c,t} : \mathbb{R}_+ \mapsto \mathbb{R}_+$$

Furthermore, we assume a demand saturation level $\overline{x}_c(t)$ beyond which customers no longer obtain any utility from additional electricity consumption. In other words,

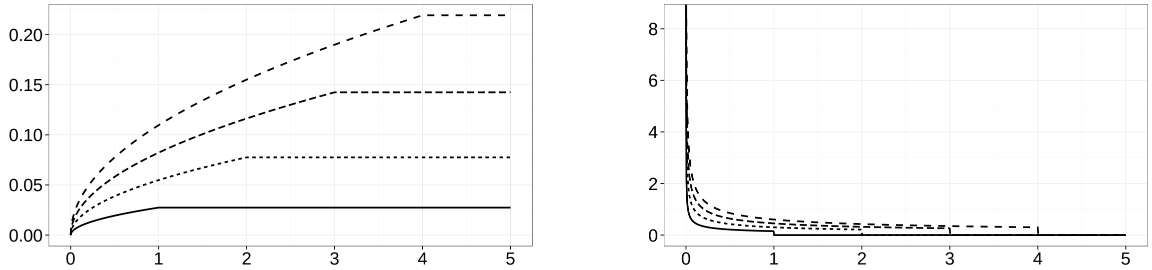
$$U_{c,t}(x_c(t)) = U_{c,t}(\overline{x}_c(t)) = \overline{U}_{c,t} \quad \forall x_c(t) \geq \overline{x}_c(t).$$

If the saturation level does (not) affect marginal utility, we refer to the utility function as saturation-level-dependent (independent). □

This customer model reflects the smart grid rationale of customers adapting consumption to current system conditions. Following standard economic theory, marginal utility of additional consumption is assumed to decrease, as most valuable use forms are activated first. At some point the customer will not be able to put additional energy allocations to any meaningful use. In our analysis, we make the following nonrestrictive assumptions: A (realistic) market will not allow orders of infinitesimal small quantities of electricity. Thus, the quantities are discretized.

Notation 14 (Discretization Granularity D): The allowed order quantities are discretized with granularity D. Consequently, D is the smallest quantity that a costumer can bid on.

The admissible $x_c(t)$-values are also discretized: $x_c(t) \in \{n \cdot D | n \in \mathbb{N}\}$. Similarly, we discretize the $\overline{x}_c(t)$ values. To model the temporal pattern of the energy-usage behavior of customers, the saturation levels $\overline{x}_c(t)$ fluctuate over time in tune with a representative energy-demand profile. When considering families of utility functions, we interpret the concavity of each function as differing



(a) Utility

(b) Marginal Utility

Figure 4.4: Illustration of utility: $\overline{x}_c(t)$ dependent

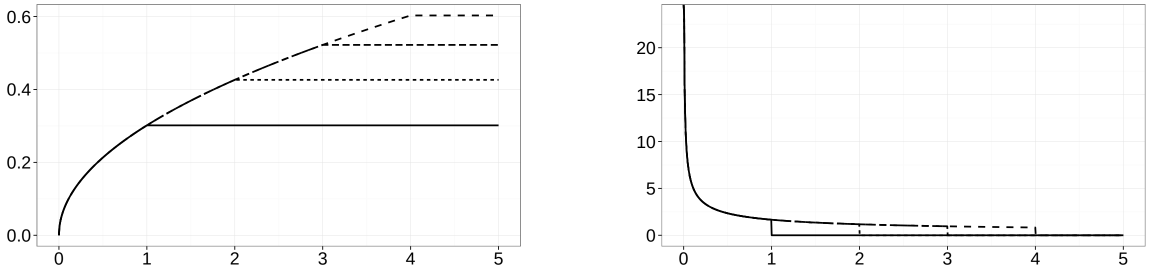
levels of load flexibility. In the case of a linear utility function, marginal utility from consumption is constant; hence load shedding has a constant cost. Conversely, for a very concave function, shedding utility losses at high load levels are limited. In the following, ϕ denotes the demand flexibility.

Notation 15 (Flexibility Level ϕ): The symbol ϕ_c denotes the demand flexibility of customer c. Higher ϕ values indicate more flexible demand.

Bidding Behavior Because $U_{c,t}$ provides a mapping from allocation to utility space, we can express a customer's optimal bidding behavior under this utility function using the marginal utility $U'_{c,t}$: Instead of placing a single price-quantity bid, a rational customer will rather place a continuum of bids with infinitesimal quantity and decreasing bid price to match his or her marginal utility function. Under our market discretization scheme, customers will place up to $n = \frac{\overline{x}_c(t)}{D}$ bids with quantity D each. The corresponding optimal bid prices are then $U'_{c,t}(i \cdot D)$ with $i = 1...n$.

We distinguish between utility functions that are increasing constantly, dependent or independent of the saturation level $\overline{x}_c(t)$. The constant utility reflects a customer having a fixed valuation for electricity. In the constant case, a customer will place a single bid with his or her constant valuation as a limit price and the saturation level as quantity. In the dependent case, the saturation level affects a customer's (marginal) utility value over the complete range of allocation quantities. In contrast, for saturation-level independency the (marginal) utility is independent of the saturation level over the interval $[0, \overline{x}_c(t)]$. Figs. 4.6, 4.4, and 4.5 illustrate examples of utility and corresponding marginal utility functions.

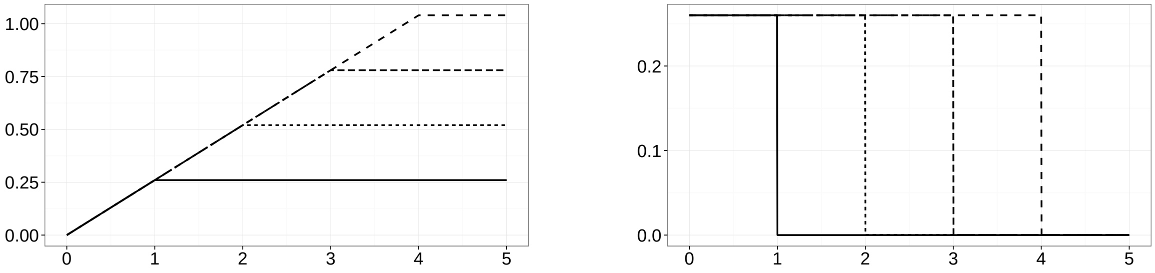
The constant utility reflects valuation in the traditional grid. In the smart grid practice, the utility of a household is a combination of saturation-level dependency and independency. The analysis of the polar cases allows us to better structure our results.



(a) Utility

(b) Marginal Utility

Figure 4.5: Illustration of utility: $\overline{x}_c(t)$ independent



(a) Utility

(b) Marginal Utility

Figure 4.6: Illustration of utility: constant increase

4.4 Model Implementation

The actual privacy costs depend on a large number of possible influence factors. This includes different privacy preferences and fluctuating supply and demand patterns. To derive meaningful results we conduct simulations based on real-world data. Additionally, simulations require an instantiations of all model components theoretically described in Section 4.3. In this section, we describe all of the details for conducting simulations.

4.4.1 Demand Model

To perform a numerical evaluation of our scenario, we need a concrete instantiation of the utility model. We study three alternatives: the constant increase of utility, one featuring dependency of marginal utility on the saturation level, and one with independence.

Let C be the constant valuation for each unit of electricity. The constant increase in utility is then defined as follows:

$$U^C(x_c(t), \overline{x}_c(t), C) = C \cdot \min\{x_c(t), \overline{x}_c(t)\}.$$

In the remaining two cases we have a parameter ϕ that represents load flexibility (i.e., concavity). To improve comparability, we normalize the utility functions with a scalar, which represents a maximum saturation level A.

Denoting the allocation by $x_c(t)$, the saturation level by $\overline{x}_c(t)$, and the flexibility level by ϕ, the function with saturation level *dependent* marginal utility (superscript D) is given by

$$U^D(x_c(t), \overline{x}_c(t), \phi) = \frac{\overline{x}_c(t)}{A} \sqrt[\phi]{\frac{\min\{x_c(t), \overline{x}_c(t)\}}{A}}.$$

For saturation-level-*independent* marginal utility (superscript I), we have

$$U^I(x_c(t), \overline{x}_c(t), \phi) = \sqrt[\phi]{\frac{\min\{x_c(t), \overline{x}_c(t)\}}{A}}.$$

By taking the first derivative with respect to $x_c(t)$ constancy, dependence and independence are easily verified: Leaving the minimum function aside, because it only reflects the upper border of the utility, the derivative of U^C consists of the constant factor C only. The derivative of U^D is dependent on $\overline{x}_c(t)$, and the derivative U^I is not.

Consumption in the simulation is based on the CER smart-metering data set (see Section 2.2.3 for more details). This data set consists of approximately 5,000 Irish homes with different numbers of inhabitants, measuring electricity consumption every 30 min over more than 1 year. For our simulations, we create a set of households in which sizes follow the distribution explained in Section 2.2.3.

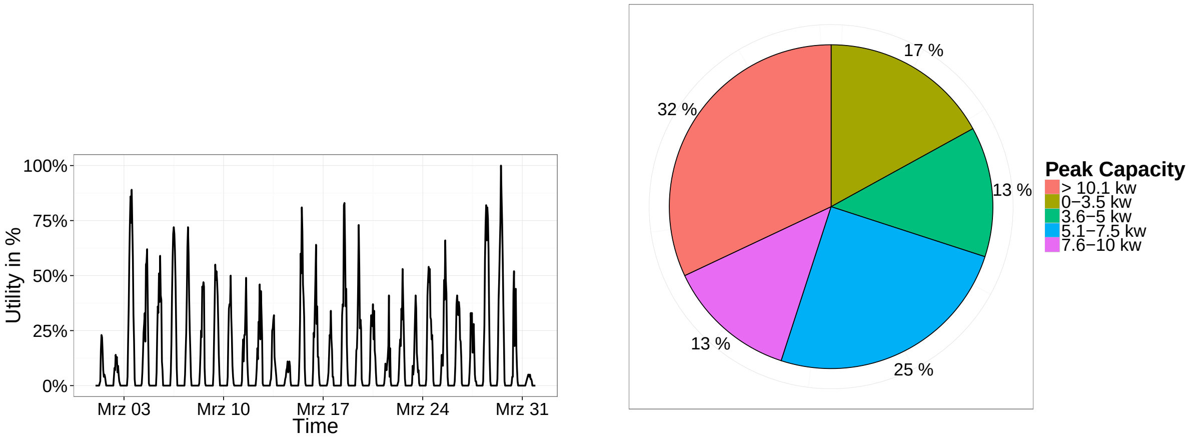
4.4.2 Market supply

We assume that there are three types of generators in the local market, namely, PV sites, CHP units, and conventional backup generation by the DSO (balancing party). PV and CHP sites are the most popular sources that feed into the low-voltage grid. We first explain our supply model and then discuss pricing schemes. Despite the tremendous amounts of technologies developed (see Section 1.1) for a 'smarter' grid, the transformation of the currently running power system has only begun [50, 84]. We evaluate how the local energy market performs in a rather *traditional* and a *smart grid–enabled* scenario. This affects the pricing schemes: The traditional model adopts prices from the current grid, whereas the smart grid–enabled model assumes different flexibility levels and strategic price determinations.

Photovoltaic Sites Supply The energy output of PV sites depends on the peak capacities, the real electricity production depending on the weather, and a spread that reflects the fact that PV sites on roofs have different angles to the sun.

The peak capacity is the maximal capacity of a PV site, and depends on the number and the quality of the solar modules installed. Figure 4.7b shows the distribution of peak capacities considered for PV sites with a range below 50 kW [109]. In our simulation, PV panel sizes are distributed according to recent German installation data censored at a maximum of 11 kWp. The used capacity is the average percentage of the peak capacity per time interval that is actually achieved under real weather conditions. To compute this, we have used our data set from the energy production of a photovoltaic site (see Fig. 4.7a). Finally, we have to consider a spread resulting from the fact that on-the-roof sites are mounted at different angles to the sun. For example, a PV site on a roof that is mounted eastward produces most energy in the morning, and a smaller amount of energy during the rest of the day. Accordingly, a site that is oriented westward produces most energy in the afternoon. We simulate the different mounting positions relative to the insulation angle by random shifts in the 'time' and 'produced energy' dimensions. We determine the random shifts with uniform distributions with parameters 0 to 0.1 for the 'produced energy' dimension, and 0 to 0.3 for the time dimension.

Combined Heat and Power Units Supply CHP sites are bound to the heating demand (CHP sites are 'heat led'). The demand depends on various parameters, e.g., the capacity of the heat storage and isolation of the house. If the heating of a house is activated, CHP sites produce a fixed amount of energy per time interval. If heating is deactivated, a CHP site consumes a small amount of energy for internal operations (see Fig. 4.8). For the sake of simplicity, we simulate a CHP site with a rectangular-shaped electricity production without consuming electricity when not operational. Generation availability is driven by heating demand and thus depends on heat storage or insulation but not on market parameters. From our data sources, we have also extracted the probability distribution of starting times and durations. We use this distribution to randomly generate start and stop times for a large set of small CHP sites with a capacity below 1 kW.



(a) PV Electricity Supply

(b) Distribution of PV Capacities

Figure 4.7: Parameters for the simulation of photovoltaic sites

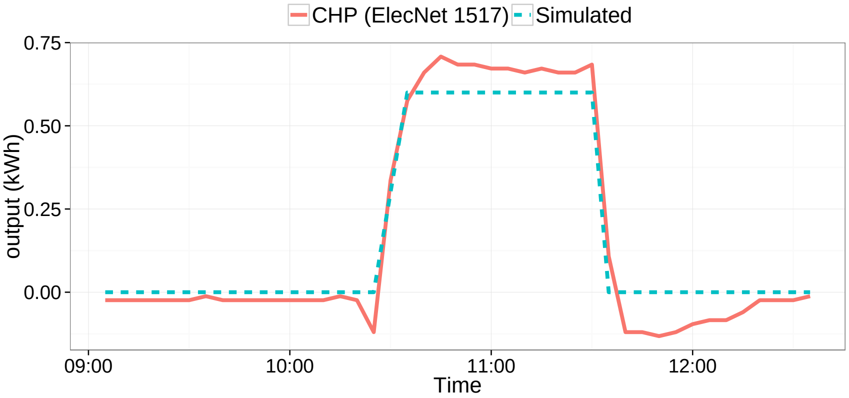


Figure 4.8: Illustration of CHP electricity supply

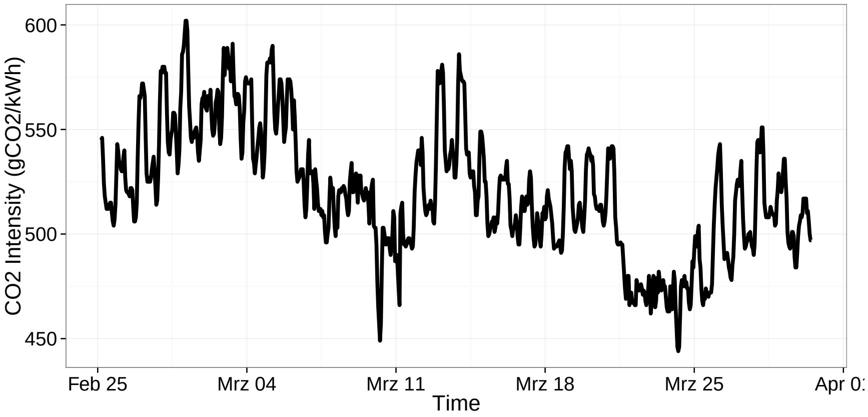


Figure 4.9: CO_2 intensity measurements

Balancing Party Supply As illustrated in the technical architecture, the DSO has access to the higher grid levels and thus is assumed to be able to provide electricity for all possible demand. However, electricity consumed from the balancing party produces CO_2 emissions. Depending on the mix of primary energy sources at a specific time of the day, intensity may vary. Our data source is the European Energy Exchange[1]. We model the emissions per kWh at time slot t as $\omega(t)$. Figure 4.9 illustrates sample data on CO_2 intensity.

Traditional Pricing Scheme *Buy Orders:* The limit price of a buy order is the upper bound for the price for energy that an consumer is willing to pay. If a local producer requires a price above the limit of the buy order, a rational consumer could simply buy his or her supply from the balancing party. Thus, the price of the balancing party is the upper bound of the market price. In 2013, German households could buy electricity from energy providers at a price of approximately $0.26\frac{€}{kWh}$ [35]. End-consumer electricity prices increased slightly in the last few years; Hence, we assume a price of $0.27\frac{€}{kWh}$ in our future market scenario. Consequently, each consumer valuates electricity with the utility function U^C and $C = 0.27\frac{€}{kWh}$.

Prices for balancing party supply: The 'traditional' energy provider charges $0.27\frac{€}{kWh}$ for consumed electricity. This is consistent with the current electricity supply with a uniform price.

The limit of a sell order is the lower bound for the compensation that a local producer demands from the consumer. In particular, PV and CHP prices are assumed to follow recent regulations.

Prices for PV supply: The compensation fee for German PV sites has been regulated in the German Renewable Energy Act [14]. As Figure 4.10 shows[2], the costs depend on the year of

[1] `http://www.eex.com` (European Energy Exchange)

[2] The German Renewable Energy Act distinguishes among PV sites with different peak capacities. However, because only small sites directly feed into low-voltage circuits, we only have to consider one price.

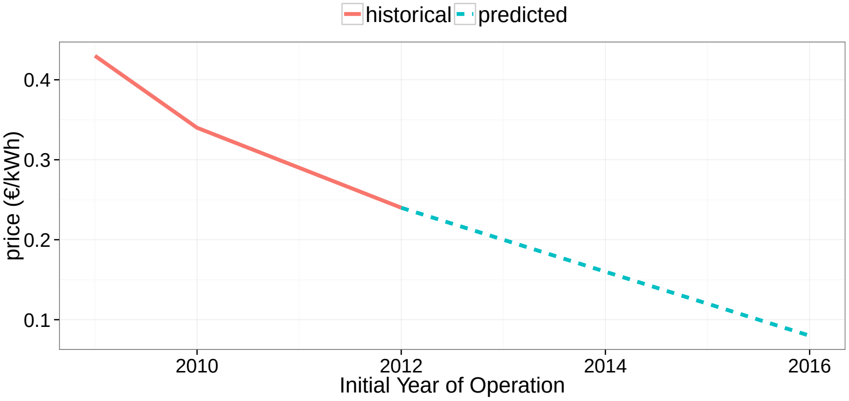


Figure 4.10: Compensation for PV sites

construction and degrade annually. The costs from 2013 to 2016 are predicted [64] (Fig. 4.10). To assign limit prices to different numbers of PV sites, we use the following model: We assume that no PV sites have been constructed before 2012, as prices would not be competitive. For each year between 2012 and 2016, one-fifth of PV sites are constructed, which offer energy with the limit prices shown in Figure 4.10.

Prices for CHP supply: In Germany, the compensation price for electricity produced by small CHP sites is fixed to $0.11\frac{€}{kWh}$ for 10 years by the CHP Act [16]. If an consumer required a lower limit, a CHP site operator could simply sell its energy to the DSO at $0.11\frac{€}{kWh}$.

Smart Grid–Enabled Pricing Scheme *Buy orders:* Limit prices of the orders follow the marginal utility expressed in U^D, respectively U^I. The utility functions require the specification of ϕ and A; we provide concrete values in our evaluation.

Prices for CHP and PV supply: The electricity output from PV sites has no marginal generation costs, consequently, a zero asking price is quoted. Under heat-led operation, all operational costs can be attributed to heating demand with electricity output arising as a byproduct. Consequently, CHP output is also bid into the local market at a zero limit price.

Prices for balancing party supply: A standard economic assumption for modeling conventional backup generation is a convex cost function [98]. This reflects the technological heterogeneity on the supply side (merit order dispatch). A quadratic cost function is a simple example of such a supply curve [103]. We follow this rationale and assume that the balancing party quotes a bid price of $p(x) = \alpha \cdot x^2$ for the xth unit of output.

4.4.3 Energy Storage

Because of increased uncertainty on the supply side, energy storage is expected to play a more important role in future smart grids. Storage operators can capitalize on the expected price fluctuations. This energy arbitrage motive has been investigated in the recent literature [98]. Our analysis adds a new economic perspective of active storage management. We investigate to what extent energy storage can mitigate the welfare loss resulting from privacy-enhancing methods in local energy markets. For the sake of generality, we assume that each customer owns a generic energy store with capacity $\overline{\mathcal{B}}_c$ (in kWh), fill level $\mathcal{B}_c(t) \in [0, ..\overline{\mathcal{B}}_c]$, and efficiency level $\mathcal{L} < 1$.[3]

Departing from an economic storage operation paradigm, we posit a simple strategy. Denoting the deviation from the current saturation level $\overline{x}_c(t)$ by ξ, the following cases are possible:

1. $\xi > 0$ — Whenever privacy enhancement results in an upward distortion, the amount $\min\{\mathcal{L} \cdot \xi, \frac{\overline{\mathcal{B}}_c - \mathcal{B}_c(t)}{\mathcal{L}}\}$ is transferred to the storage unit; and
2. $\xi < 0$ — In case of an allocation shortfall resulting from high market prices or a downward distortion, customers withdraw the amount $\min\{\xi, \mathcal{B}_c(t)\}$ from the energy store.

This policy could be improved, e.g., by adopting dynamic threshold levels. However, by focusing on this rather naïve policy, we can isolate interactions between privacy enhancement and the presence of storage capacities.

4.5 Conclusions

The proposed local energy market is a meaningful application-specific data-quality measure for the effects of privacy-enhancement methods on time series. The provided measures are comprehensible: Welfare, respectively relative welfare, reflects the savings of individuals when participating in the market. Saved CO_2 amounts can be compared with other efforts for reducing emissions. The local energy market is a popular application that is investigated in the smart grid context. Thus, the results are discriminative at least for this specific application. However, welfare loss and emissions depend on typical effects in the smart grid: For example, allocative inefficiency resulting from privacy enhancement has a stronger impact on the measures the more renewable energy is available, or the higher the CO_2 intensity is. Because of the nature of renewable sources, more electricity is available during the day and less during the night. Consequently, perturbation of consumption values during the night have a different effect on the market results than during the day. This suggests, that the results of the local energy market are also discriminative for other smart grid applications. Configurations of supply or demand could also be altered to reflect the behavior of the application in question, e.g., changing peak hours of the supply or demand side.

[3] For the sake of exposition, we only account for losses during charge. Furthermore, we do not explicitly consider indirect storage options (e.g., hot water storage, electric vehicles). Such alternative storage systems would not alter the general results obtained.

However, it is still questionable how privacy-enhancement methods such as PACTS or others actually affect market performance or emissions. In addition, applying privacy-enhancement methods may also influence theoretical properties of such markets. We evaluate both in the following chapter.

Chapter 5

Impact of Privacy Enhancement on Electronic Markets

In the previous chapter, we have seen that the local energy market scenario, as an electronic market, is a comprehensible and discriminative utility measure for privacy-enhancement methods. The resulting welfare and CO_2 emissions rate the decrease of data quality induced by the data modifications of privacy-enhancement methods. The measure exploits the fact that privacy-enhancement methods reduce allocative efficiency of such markets. In turn, this gives insight into the question of whether electronic markets are still effective under the presence of such methods. We contribute a detailed analysis of different privacy-enhancement approaches applied to local energy markets in real-world scenarios (**C.4**). This includes the definition of general properties of privacy-enhancement methods. Furthermore, we show that a privacy-aware auction retains incentive compatibility with respect to valuations if the privacy-enhancement method is monotonic and marginal utility is independent of the demand level.

Because of the large number of possible influence factors (e.g., customer privacy preferences, demand-side flexibility, and supply and demand patterns), it is difficult to fully characterize the welfare loss an saved emissions in a general fashion. For instance, realistic supply and demand patterns are complex random processes, and the applied privacy-enhancement methods add complexity as well. A general model covering all of these details would lack expressiveness. Therefore, we instantiate a numerical evaluation using empirical load and generation data. Using simulations, we quantify the costs and the emissions of privacy enhancement. Specifically, we assess the economic effect of varying numbers and types of generators, demand properties, and storage endowments. The experiments illustrate the relationship between privacy enhancement and welfare loss, respectively emission increase. Furthermore, we can quantify the positive effect of storage in the presence of privacy-enhancement methods. Small-scale electricity storage can reduce privacy-induced welfare loss by 70%. Our findings underline the validity of the proposed utility measure, as the effects of privacy enhancement are controllable and maintain important market properties like incentive compatibility. In particular, we provide the following:

- We determine formal characteristics of the interplay among components of our model. This includes the the definition of general properties of privacy-enhancement methods;

- We show that privacy-aware auctions retain important properties like incentive compatibility on a theoretical basis; and

- In our numerical evaluation we quantify the costs of privacy enhancement. We assess the economic effects of varying numbers and types of generators, demand properties, and storage endowments. The numerical evaluation also quantifies the effects of privacy enhancement on CO_2 emissions.

The numerical analysis complements the utility measure introduced in Chapter 4 with a comparison of different privacy-enhancing methods and parameters. The remainder of the chapter is structured as follows: First, we need to define the general properties of privacy-enhancement methods (Section 5.1) before proving theoretical properties on the impact on electronic markets (Section 5.2). We complement the theoretical results with a numerical evaluation (Section 5.3) and conclude (Section 5.4). The contents of this chapter are published in [12] and [57].

5.1 Properties of Privacy-Enhancement Methods

A privacy mechanism $\mathcal{M}$ takes a data set, e.g., a set of time series $\mathcal{F}$ and parameters p, and returns a privacy-enhanced representation, i.e., $\mathcal{F}' = \mathcal{M}_p(\mathcal{F})$. In our smart grid scenario, the time series are given by consumers' saturation levels: $\mathcal{F} = \{\overline{x}_c(t) | t \in \mathcal{T}, c \in \mathcal{C}\}$. The parameter p is a method-specific parameter that determines the level of privacy achieved. The privacy-enhancement method modifies the values of the time series, in our case the saturation level values: $\mathcal{F}' = \mathcal{M}(\mathcal{F}) = \{\widetilde{\overline{x}}_c(t) | t \in \mathcal{T}, c \in \mathcal{C}\}$. Various methods exist with different approaches for preserving privacy in a set of time series (see Section 3.1.2 and PACTS in Section 3.2)). To derive theoretical results independent of an actual method, we need to come up with the general properties of privacy-enhancing methods.

A central distinction is the one between *deterministic* and *randomized* privacy-enhancement methods.

Definition 38 (Deterministic): A privacy-enhancement method is deterministic if the results of several runs are the same with the same input: $\mathcal{F}'_1 = \mathcal{M}_p(\mathcal{F}) \wedge \mathcal{F}'_2 = \mathcal{M}_p(\mathcal{F}) \Rightarrow \mathcal{F}'_1 = \mathcal{F}'_2$. □

Definition 39 (Randomized): The privacy-enhancement method depends on random calculations, and this may lead to different results if the method is run several times. The probability that the method returns a certain privacy-enhanced set $\mathcal{F}'$, $P(\mathcal{M}_p(\mathcal{F}) = \mathcal{F}')$, is the same for each run. □

The rationale behind the following notions is to further characterize the effect of different privacy-enhancement methods.

Definition 40 (Balanced Modifier): Let $\mathcal{F}' = \mathcal{M}(\mathcal{F})$. $\mathcal{M}$ is a balanced modifier, if the following holds for all consumers $c \in \mathcal{C}$:

$$\sum_{\forall t \in \mathcal{T}} \widetilde{x}_c(t) - \overline{x}_c(t) = 0$$

A randomized privacy-enhancement method is a balanced modifier if the expected value of the sum of these differences equals zero. □

Definition 41 (∪-homomorphism): Assume that $\mathcal{F}_1$ and $\mathcal{F}_2$ is an arbitrary partitioning of the time series in $\mathcal{F}$:

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \wedge \mathcal{F}_1 \cap \mathcal{F}_2 = \emptyset$$

A deterministic privacy-enhancement method is a homomorphism of ∪ if the following holds:

$$\mathcal{M}(\mathcal{F}_1) \cup \mathcal{M}(\mathcal{F}_2) = \mathcal{M}(\mathcal{F})$$

A randomized privacy-enhancement method is a homomorphism of ∪ if the following holds:

$$P(\mathcal{M}(\mathcal{F}_1) \cup \mathcal{M}(\mathcal{F}_2) = \mathcal{F}') = P(\mathcal{M}(\mathcal{F}) = \mathcal{F}')$$

□

In the following, we show that a privacy-enhancement method, which is a homomorphism of ∪, will modify the time series independently of each other.

Lemma 6: *Let $\mathcal{F}' = \mathcal{M}(\mathcal{F})$. If $\mathcal{M}$ is a ∪-homomorphism, the modifications of time series $\widetilde{x}_c(t) \in \mathcal{F}'$ are independent of the time series of any other consumer $c' \neq c : \overline{x}_{c'}(t)$*

Proof: Let $\mathcal{F}_1 = \{\overline{x}_c(t)\}$ and $\mathcal{F}_2 = \mathcal{F} \backslash \mathcal{F}_1$. By definition, $\mathcal{F}_1$ and $\mathcal{F}_2$ are partitions of $\mathcal{F}$. Because $\mathcal{M}$ is a ∪-homomorphism, $\mathcal{M}(\mathcal{F}_1) \cup \mathcal{M}(\mathcal{F}_2)$ equals $\mathcal{M}(\mathcal{F})$. In particular, the resulting $\widetilde{x}_c(t)$ is independent of possible other time series in $\mathcal{F}$. □

For instance, k-anonymity [108] usually does not have this property: The output of most implementations depends on the groups created. In turn, adding symmetric random noise is a ∪-homomorphism.

The privacy parameters p influence the privacy enhancement. In the following we define an order.

Definition 42 (Order of Privacy Parameters ($p_1 > p_2$): Let p_1 and p_2 be different parameters for privacy method $\mathcal{M}_p$. p_1 is greater than p_2 if $\mathcal{M}_{p_1}(\mathcal{F})$ provides a better privacy protection than $\mathcal{M}_{p_2}(\mathcal{F})$ in terms of the definition of $\mathcal{M}_p$. □

Commonly known distance metrics, e.g., the L1-Norm, quantify the distance between two time series. Choosing a greater privacy parameter may lead to a larger distance if the privacy method is monotonically increasing, as defined in the following. Let $dist(\overline{x}_c(t), \widetilde{x}_c(t))$ be such a distance metric.

Definition 43 (Monotonically Increasing): Let p_1, p_2 be two privacy parameter choices for $\mathcal{M}_p$ with p_1 having greater order than p_2, that is $p_1 > p_2$. Furthermore, $\widetilde{x}_c^1(t) \in \mathcal{M}_{p_1}(\mathcal{F})$ and $\widetilde{x}_c^2(t) \in \mathcal{M}_{p_2}(\mathcal{F})$ are time series obtained by applying $\mathcal{M}_p$ on the same time series $\overline{x}_c(t) \in \mathcal{F}$.

A *deterministic* privacy-enhancement method is monotonically increasing with respect to a metric $dist(\cdot)$, if the following holds for:

$$dist(\overline{x}_c(t), \widetilde{x}_c^1(t)) \geq dist(\overline{x}_c(t), \widetilde{x}_c^2(t)).$$

A *random* privacy-enhancement method is monotonically increasing with respect to a metric $dist(\cdot)$ if in expectation the following holds:

$$\mathbb{E}\left[dist(\overline{x}_c(t), \widetilde{x}_c^1(t)) - dist(\overline{x}_c(t), \widetilde{x}_c^2(t)\right] \geq 0$$

□

Intuitively, a privacy enhancement method is monotonically increasing if greater privacy-parameter choices give rise to greater changes to the original time series values.

5.2 Theoretical Results

We now derive formal results on the impact of privacy enhancement on local energy markets. In the following we assume that the time series are of infinite length. We also assume non triviality of the privacy-enhancement methods, i.e., we exclude the case that $\widetilde{x}_c(t) = \overline{x}_c(t), \forall t \in \mathcal{T}$.

Lemma 7: *The welfare loss is monotonically increasing for greater privacy parameters if the privacy-enhancement method is monotonically increasing and a balanced modifier.*

Proof: Assume $\mathcal{F}' = \mathcal{M}_p(\mathcal{F})$. Further, let d be the difference between the saturation level and the privacy-enhanced saturation level of consumer c on time slot t: $d = \overline{x}_t^c(-)\widetilde{x}_c(t)$. If $d > 0$, the higher d the lower the $\widetilde{x}_c(t)$ and potentially the higher the welfare loss. c may not get electricity allocated even if the marginal utility is greater than zero because there are no bids exceeding $\widetilde{x}_c(t)$. A similar result holds for $d < 0$: The smaller d the higher the potential welfare loss, because c may allocate energy at a price greater than zero, even if the marginal utility is zero. Depending on the actual utility functions, the welfare loss is higher for $d > 0$ or $d < 0$. However, if the privacy-enhancement method is a balanced modifier, the sum of the welfare loss of all time slots remains the same. Let the privacy-enhancement method $\mathcal{M}_p^1$ be monotonically increasing, then the welfare loss for more restrictive privacy requirements is higher, as the distance between $\overline{x}_c(t)$ and $\widetilde{x}_c(t)$ also increases for $p_1 > p$. □

Lemma 8: *In the presence of storage, the welfare loss is equal or smaller than without storage if the privacy-enhancement method is a balanced modifier.*

Proof: Let $\mathcal{M}_p$ be a balanced modifier and $\mathcal{F}' = \mathcal{M}_p(\mathcal{F})$. Assume that there exists a $t_1 \in \mathcal{T}$ where $\widetilde{x}_c(t_1) > \overline{x}_c(t_1)$. Because $\mathcal{M}_p$ is a balanced modifier, we assume that there exists a t_2 where $\widetilde{x}_c(t_2) < \overline{x}_c(t_2)$. If the allocation at t_1 is also greater than $\overline{x}_c(t_1)$, the additional electricity bought is stored and used in times of undersupply, or at t_2. If the storage did not exist, the additional electricity bought at t_1 would not return in utility. Only in the case that after t_1 there is no time slot with undersupply, storage cannot reduce the welfare loss. □

Lemma 9: *Privacy-induced welfare loss is weakly decreasing in demand flexibility if the utility function is saturation-level dependent.*

Proof: Assume a privacy-enhanced market allocation for a given flexibility level ϕ. If the demand flexibility is raised to $\phi' > \phi$, the assumed concavity of the utility functions leads to the following effect: The marginal utility $U'_{c,t}(i \cdot D)$ with $i = 1...n$ for small is is higher for ϕ' than for ϕ, and drops faster for greater is. Formally, let $\left[U'_{c,t}(i \cdot D)\right]_\phi$ be the marginal utility for flexibility level ϕ, then there exists a threshold $\widehat{i}$ where

$$\left[U'_{c,t}(\widehat{i} \cdot D)\right]_\phi \leq \left[U'_{c,t}(\widehat{i} \cdot D)\right]_{\phi'}$$

and

$$\left[U'_{c,t}((\widehat{i}+1) \cdot D)\right]_\phi > \left[U'_{c,t}((\widehat{i}+1) \cdot D)\right]_{\phi'}$$

holds. Because the higher valued units have a higher probability of being allocated, and a lower probability of being omitted if the privacy-enhancement method changes the saturation level, the welfare loss is weakly lower for ϕ'. □

If the utility is independent of the saturation level, the actual saturation $\overline{x}_c(t)$, respectively $\widetilde{x}_c(t)$, does not necessarily reach the threshold $\widehat{i}$. Thus, Lemma 9 does not hold for saturation-level independent utility.

A privacy-enhancement method leads to a distortion of saturation levels, which has the following effect: Replacing the saturation level $\overline{x}_c(t)$ with a distorted value $\widetilde{x}_c(t)$ could naturally have a *quantity* effect on the resulting bidding behavior. This becomes evident in Definition 37: Inflated values, i.e., $\widetilde{x}_c(t) > \overline{x}_c(t)$, lead to positive marginal utility assessments when the marginal utility is zero in the nondistorted case. Discounted values, in turn, i.e., $\widetilde{x}_c(t) < \overline{x}_c(t)$, yield premature zero-marginal-utility assessments. However, remember that we ruled out untrue saturation-level reports to the privacy-enhancement method $\mathcal{M}$, as this potentially leads to a privacy breach. The semantic of $\mathcal{M}$ is defined on sensitive and true personal data; the effects if applied to untrue data are unknown. Furthermore, a deviation from the true saturation level will not influence the bidding quantities of others if the privacy-enhancement method is a $\cup$-homomorphism.

Although we rule out quantity misreports, we are interested in characterizing privacy-aware markets that induce consumers to reveal their true valuation.

Definition 44 (Incentive Compatibility): A privacy-aware market mechanism is (in expectation) incentive-compatible with respect to valuation if consumers cannot (in expectation) profitably deviate from placing bids that reflect their true valuation. □

Thus, consumers will bid according to their true valuation in the presence of an incentive-compatible privacy-enhancement method.

Lemma 10: *An incentive-compatible market mechanism retains this property in the presence of privacy enhancement, if the following holds: The privacy-enhancement method is an* $\cup$*-homomorphism and the utility function is saturation-level independent.*

Proof: Although the distortion of the saturation level always affects the optimal quantity, it does not necessarily have an effect on the optimal bid price. The occurrence of a price effect hinges on the structure of the customer-utility function: If $U'_{c,t}$ is independent of $\overline{x}_c(t)$, the bid price will always reflect the customer's true valuation for all demand increments $x \in [0, \min\{\overline{x}_c(t), \widetilde{x}_c(t)\}]$. In contrast, if the utility function is saturation-level dependent, this leads to a price effect, and the consumers may strategize and report prices that are different from their true valuation. The $\cup$-homomorphism property excludes incentives from true reports, as other consumers are not influenced. Consequently, incentive compatibility is preserved if the utility is independent from the saturation level, and the privacy-enhancement method is a $\cup$-homomorphism. □

5.3 Numerical Evaluation

Because a purely theoretical analysis would lack expressiveness, we conducted a number of experiments with different privacy-enhancement methods and real-world data. We describe the results in the following.

5.3.1 Privacy-Enhancement Methods Considered

We conduct experiments with four privacy-enhancing methods from three different classes: A k-anonymity derivative on the time series [86] (see Section 3.1.2) as a representative of 'anonymization'; a 'perturbation' approach we call wavelet privacy [88] (see Section 3.1.2), and a slightly modified version of this algorithm, which retains incentive compatibility, called 'IC wavelet privacy'. PACTS (see Section 3.2) provides 'provable privacy' for the time series. In the following, we investigate which properties the proposed algorithms have.

We assume the saturation levels $\overline{x}_c$ as input time series. The privacy-enhancement mechanisms then return the modified saturation levels $\widetilde{x}_c$ for the privacy-enhanced bidding. Refer to Section 4.3.2 for a detailed description on the integration of privacy-enhancement methods in the market.

k-anonymity We provide a pseudo code implementation in Algorithm 1. This method is *randomized*, as time series are randomly selected yielding to possible different results for each run. The method replaces time-series values $f_p(t)$ with the average of all k time series belonging to the same group. By definition, the differences between $f_p(t)$ and $f'_p(t)$ add up to zero, making this method a *balanced modifier*. Partitioning the data set into two distinct groups and applying the method to both obviously yields a different result than modifying the whole set. For example, think of a group of k time series in $\mathcal{F}$ having small distances. Considering the whole data set, such time

Property	**k-Anonymity**	**Wavelet Privacy**	**IC Wavelet Privacy**	**PACTS**
Deterministic				
Randomized	✓	✓	✓	✓
Balanced Modifier	✓	✓	✓	(✓)
$\cup$-*homomorphism*		✓	✓	✓
Monotonically Increasing	✓		✓	(✓)

Table 5.1: Property overview for privacy-enhancing methods considered

series will most likely form a k group. If these time series are partitioned into two different sets, the resulting k-anonymous data set will differ: k-anonymity is not a $\cup$-*homomorphism*. The order of privacy parameter k is canonical. In terms of k-anonymity, a higher k provides better protection of privacy. Assume that $k_1 > k_2$ holds, a data set that is k_1-anonymous requires more modifications than for k_2. Consequently, this method is *monotonically increasing*.

Wavelet Privacy-Enhancement Algorithm Refer to Section 3.1.2, respectively Algorithm 2, for a detailed explanation of this privacy method. The perturbation is based on noise and is obviously *randomized*. Because all time series are treated independently, it is also a $\cup$-*homomorphism*. Finally, the symmetry of the noise distribution ensures that the method is a *balanced modifier*.

However, the method is not monotonically increasing: Applying a higher threshold $\sigma_1 > \sigma_2$ most likely results in noise with a higher variance, but is only applied to fewer coefficients. In general, we cannot assess whether $\mathcal{M}_{\sigma_1}(f_p)$ will distort a single data point to a larger extent than $\mathcal{M}_{\sigma_2}(f_p)$.

Incentive Compatible Wavelet Privacy-Enhancement Algorithm In what follows, we propose a modification of the wavelet privacy-enhancement algorithm, referred to as *incentive-compatible wavelet privacy-enhancement algorithm (IC-wavelet privacy)*. As shown in Lemma 10, a privacy-enhancement algorithm needs to be monotonously increasing to retain in-expectation incentive compatibility with respect to valuation. Our modification achieves monotonicity by decoupling the threshold for coefficients and the noise variance. To this end, we introduce a parameter ξ that determines the standard deviation of the applied noise. The detailed implementation is given in Algorithm 5. For a fixed σ, the modified algorithm is monotonically increasing in ξ: Choosing $\xi_1 > \xi_2$ leads to a higher expected distance between original and perturbed time series compared with ξ_2. Thus, it fulfills all requirements of Lemma 10. By setting $\xi = \sigma$, the modified algorithm is identical to the unmodified wavelet privacy-enhancement algorithm (see Section 3.1.2).

Choosing σ and ξ: The same (absolute) σ may have very different effects on two different time series: $\overline{x}_{c_1}(t)$ and $\overline{x}_{c_2}(t)$. σ may lead to a lot of modified coefficients in $\widetilde{x}_{c_1}(t)$ because they exceed σ, whereas $\widetilde{x}_{c_2}(t)$ remains unmodified. To keep the parameters comparable, we choose σ and ξ relative to the standard deviation of the currently modified time series. Let $\sigma, \xi \in [0, 1]$. Then the

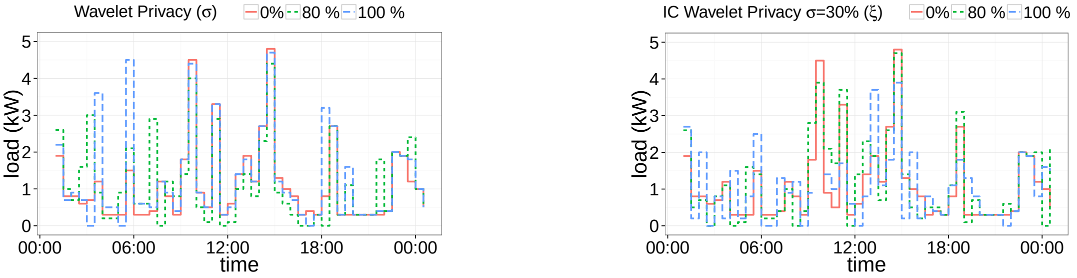
Algorithm 5: $\mathcal{M}_{\sigma,\xi}$ Modified Privacy-Enhancement Methods Corresponding to Algorithm 2

```
Input: Privacy Parameter $\sigma$
Input: Noise parameter $\xi$
Input: Set of time series $\mathcal{F}$
Result: Privacy-enhanced time series $\mathcal{F}'$
foreach $f_p(t) \in \mathcal{F}$ do
    $\widetilde{f_p(l,t)} = DWT(f_p(t))$ //Wavelet transform;
    $I_l = \left\{t : \left|\widetilde{f_p(l,t)}\right| \geq \sigma\right\}$;
    foreach level $l$ do
        $K = \sum_l K_l$ //coeffs exceeding $l$;
        $p = |N|/K$ //Noise ‘density’, N is number of coefficients;
        foreach detail $\widetilde{f_p(l,t)}$ do
            if $t \in I_l$ then
                $\widetilde{f_p(l,t)} + = GaussRnd(0, \xi\sqrt{p})$;
            end
        end
    end
    $f'_t(=)InvDWT(\widetilde{f_p(l,t)})$;
    $\mathcal{F}' = \mathcal{F}' \cup \{f'_t()\}$;
end
return $\mathcal{F}'$;
```

actual parameters σ_c and ξ_c for time series $\overline{x}_c(t)$ are the product of σ, ξ and the standard deviation of the time series $\overline{x}_c(t)$.

Figure 5.1 illustrates the effect of the privacy-enhancing techniques. The upper panel illustrates that the wavelet privacy-enhancement method is not monotonically increasing: At many points of time, the perturbed time series with $\sigma = 80\%$ has a higher distance to the nonperturbed time series than the one perturbed with $\sigma = 100\%$. In contrast, the lower panel shows the corresponding results from the monotonically increasing and incentive-compatible algorithm. The time series perturbed with the higher $\xi = 100\%$ usually has a higher distance than the one with a lower $\xi = 80\%$.

PACTS Provable privacy is achieved by applying noise to time series. This makes PACTS a *randomized* method. The noise added is symmetric, but applied to an abstracted representation. If the transformation, respectively the inverse transformation, processes noise regardless of whether it is less than or greater than zero, PACTS is also a balanced modifier. One of the design goals of PACTS is that it handles time series in isolation. Consequently, it is a $\cup$-*homomorphism*. PACTS takes ϵ and discriminative pairs of secrets $\mathcal{S}_{pairs}$ as parameters. Different sets of discriminative pairs



(a) Wavelet Privacy

(b) IC Wavelet Privacy

Figure 5.1: Examples of privacy-enhancement method realizations

are not comparable. Considering a fixed $\mathcal{S}_{pairs}$, the privacy mechanism is *monotonically increasing* in ϵ. The greater ϵ, the larger the distribution of the noise applied to the same coefficients.

5.3.2 Common Parameters

For the evaluation we chose 1,000 persons in total living in 314 randomly chosen households of the CER data set (see Section 2.2.3). The supply side is modeled as combinations of 100 or 250 PV and CHP sites.

For the traditional pricing scheme, we choose the prices proposed in Section 4.4.2: A consumer orders with a limit price of $0.27\frac{€}{kWh}$, the CHP sells electricity for $0.11\frac{€}{kWh}$, and the PV sites sell for $0.08\frac{€}{kWh}$ up to $0.24\frac{€}{kWh}$, depending on the initial year of operation. The DSO offers energy at a constant price of $0.27\frac{€}{kWh}$, and therefore does not lead to any welfare.

In the smart grid pricing scheme, we assume homogeneous utility functions ($\phi = 2$, $A = 11\ kW$) across consumers. Additionally, to quantify the effect of demand-side flexibility, we consider $\phi = 3$ and $\phi = 1$. The balancing party is parametrized with $\alpha = 2$.

In the scenarios with storage systems, we consider storage sizes of $\overline{\mathcal{B}} \in \{2.5\ kWh, 5\ kWh\}$ in line with currently marketed products. We assume storage efficiency of 80%. The time span for each simulation run is one day. We tested longer simulation horizons as well, but these results did not exhibit any substantial differences to the ones described in the following. Each experiment is repeated 10 times.

5.3.3 Privacy Parameters

For the k-anonymity we choose $k \in \{0, 2, 5, 10, 20\}$. We apply the wavelet privacy-enhancement method with $\epsilon \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For the incentive compatible version, we fix $\sigma = 30\%$ and vary $\xi \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For evaluating the impact of PACTS, we took all of the hidden sample secrets to hinder INDiC as well as re-identification. We hid whether 'Light', 'Refrigerator'

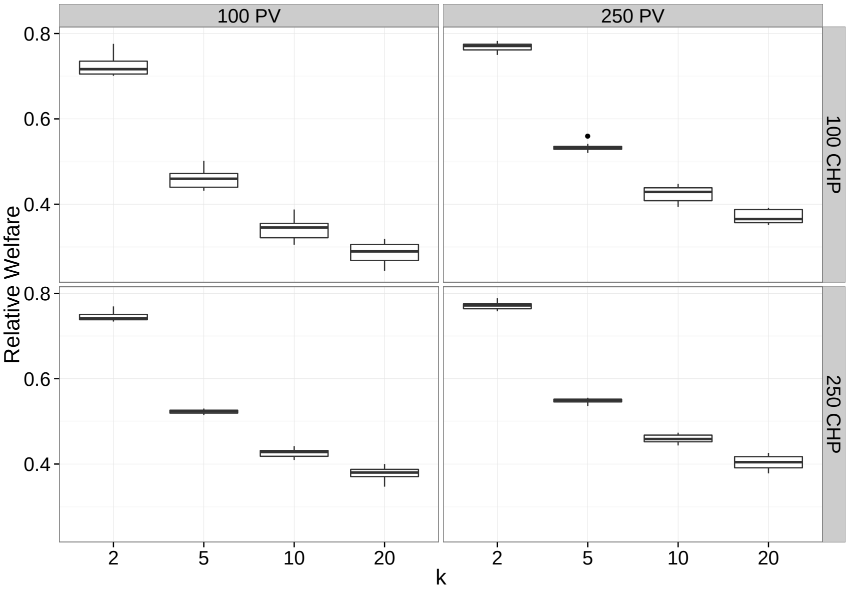


Figure 5.2: Relative welfare for traditional pricing scheme and k-anonymity

or ‘Microwave’ is in State 2 or 3 (see Table 3.3 for the specific parameters). Furthermore, we hid secrets specified for the ‘Overall’, ‘Maximum’ and ‘Minimum Consumption’ and secrets for ‘Average Wakeup Hours’ and ‘Bedtime Hours’ as defined in Section 3.4.3.

For the explanation of numerical results, we distinguish among approaches with numeric parameters (k-anonymity, wavelet privacy, IC wavelet privacy) and PACTS a privacy mechanism that requires different, nonnumeric secrets. The results of methods with numerical parameters can be set into relation to each other, i.e., parameters have an order. With PACTS instead hiding different secrets, this leads to a different and not necessarily comparable privacy results.

5.3.4 Numerical Results: k-Anonymity, Wavelet Privacy, IC Wavelet Privacy

We investigate the effect of market structure, storage endowments, and demand-side flexibility. Finally, we compare all three methods.

Effect of Market Structure First, we investigate whether and to what extent the number of PV and CHP sites influences the impact of different privacy-enhancement methods and levels to relative welfare and relative-saved emissions. This sheds light on how privacy enhancement interacts with different market structures. Figures 5.2–5.10 contain results regarding the relative welfare for the different privacy-enhancement methods, and Figures 5.11–5.19 the saved emissions.

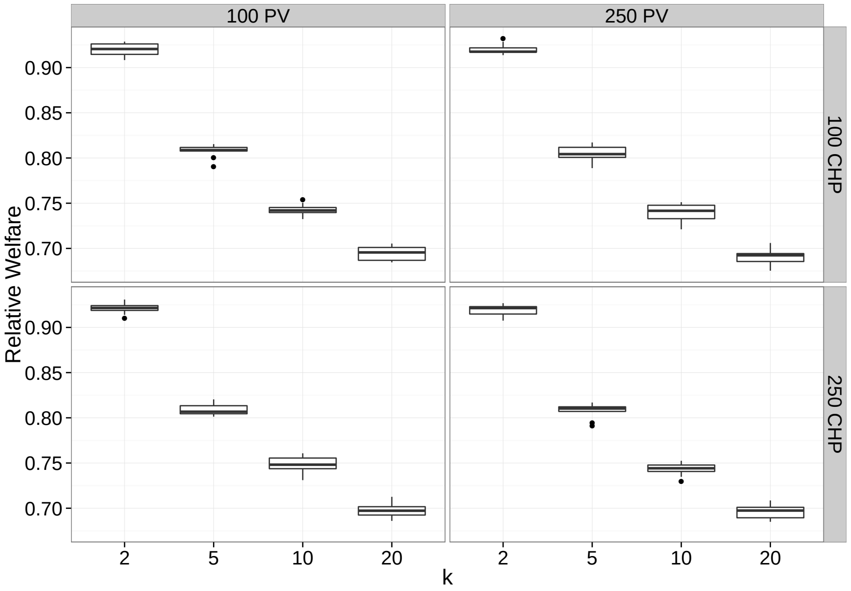


Figure 5.3: Relative welfare for the smart grid pricing scheme with saturation-level-dependent utility and k-anonymity

R.19 *Market configuration has little impact on the welfare loss as well as on the saved emissions.* Although the variance of the results (size of boxes in Figs. 5.2–5.10, respectively Figs. 5.11–5.19) is naturally higher in smaller markets with few generators, the median results are hardly affected. This holds for all privacy-enhancement methods investigated. Privacy enhancement has a similar (relative) effect, independent of the actual market structure.

This result is beneficial for privacy enhancement in such scenarios in the following way: The negative effect of privacy-enhancement methods is predictable and does not require knowledge about the current or future market structure. In turn, increasing privacy parameters has a different effect.

R.20 *Smart grid pricing scheme: Saturation-level-independent utility exhibits decreasing marginal welfare cost of the privacy level.* Although the welfare loss is strictly increasing in the privacy level for both utility specifications, saturation-independent utility exhibits decreasing marginal losses in our results. For saturation-level-dependent utility we observe almost linear behavior. This is the result of different effects of perturbed saturation levels. (Figures 5.4, 5.7 and 5.10 contain the results for the saturation-independent utility and Figures 5.3, 5.6 and 5.9 the results for the dependent utility.)

However, different utility ratings do influence CO_2 emissions, but not as strongly as the welfare.

R.21 *Smart grid pricing scheme: Dependence on utility leads to less saved emissions with privacy*

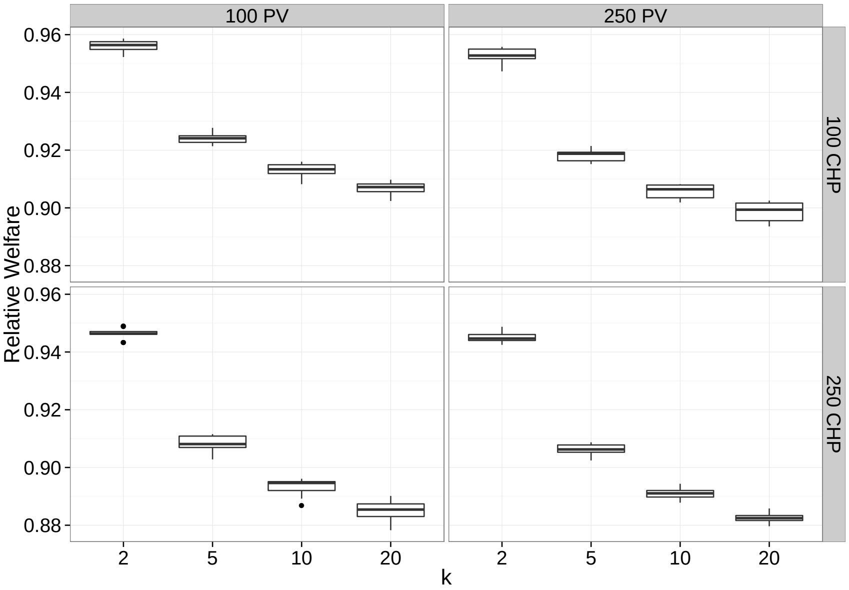


Figure 5.4: Relative welfare for the smart grid pricing scheme with saturation-level-independent utility and k-anonymity

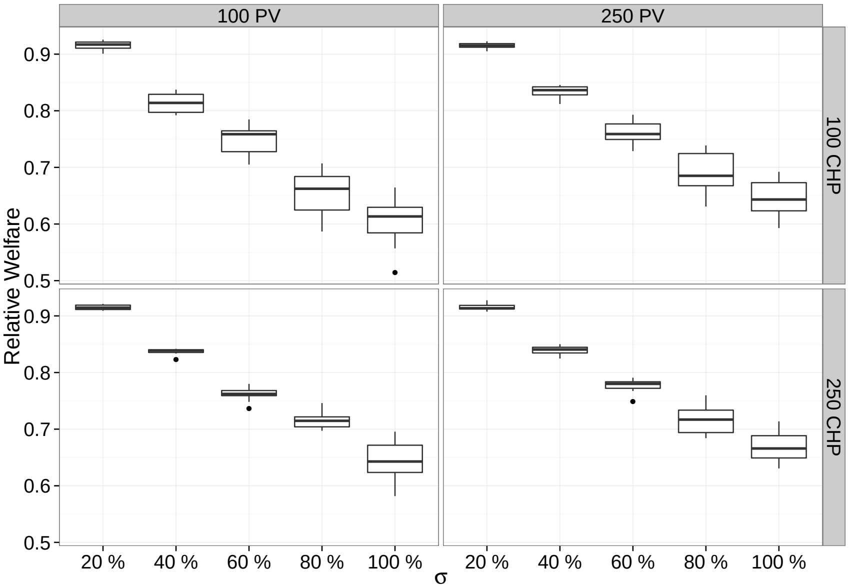


Figure 5.5: Relative welfare $\mathcal{W}'$ for traditional pricing scheme and wavelet privacy

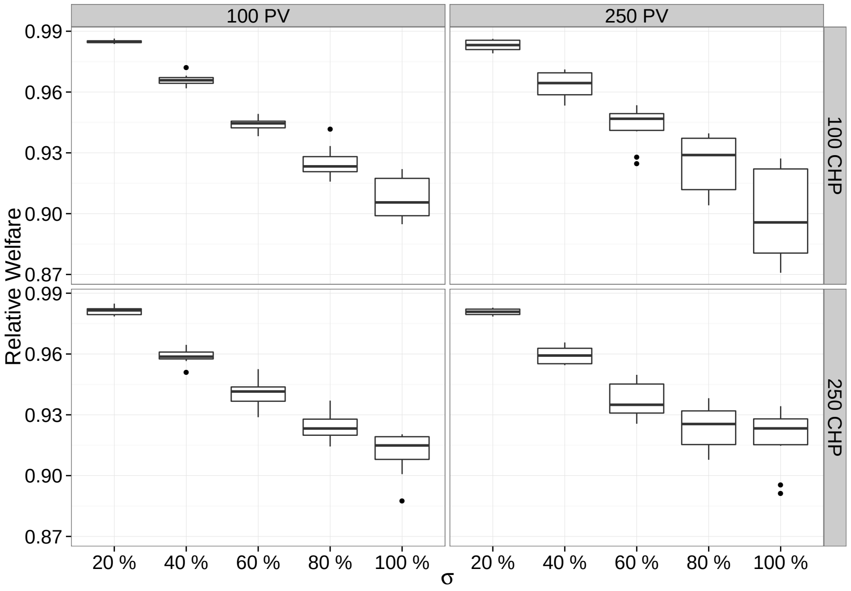


Figure 5.6: Relative welfare $\mathcal{W}'$ for the smart grid pricing scheme with saturation-level-dependent utility and wavelet privacy

enhancement. On average, the saved emissions with the saturation-level-independent utility are a little higher than for the dependent utility. Because of privacy enhancement, the limit prices of consumers also changes in the case of dependent utility. This influences the trades with the balancing party more, compared with the saturation-independent case. For the wavelet and the IC wavelet privacy, the difference between relative saved emissions is below 5%, and for the k-anonymity privacy enhancement this difference is even smaller. In all cases the negative effects are lower as on relative welfare (Figures 5.12 and 5.13 contain the results for k-anonymity, Figures 5.15 and 5.16 contain the results for wavelet privacy, 5.18 and 5.19 contain the results for IC wavelet privacy).

Comparing the traditional pricing scheme with the smart grid pricing scheme leads to more significant results.

R.22 *Privacy enhancement and the traditional pricing scheme leads to higher loss of welfare and reduces the saved emissions more compared with the smart grid pricing schemes.* In the traditional pricing scheme, each unit of electricity has the same valuation for a consumer. In turn, in the smart grid pricing scheme, valuations differ. Inaccuracies resulting from privacy enhancement usually affect units with lower valuations in the smart grid pricing scheme. Thus, welfare decreases more in the traditional scheme. Emissions depend on the amount of electricity bought from the balancing party. In the smart grid pricing scheme, the limit price of the balancing party is quadratic

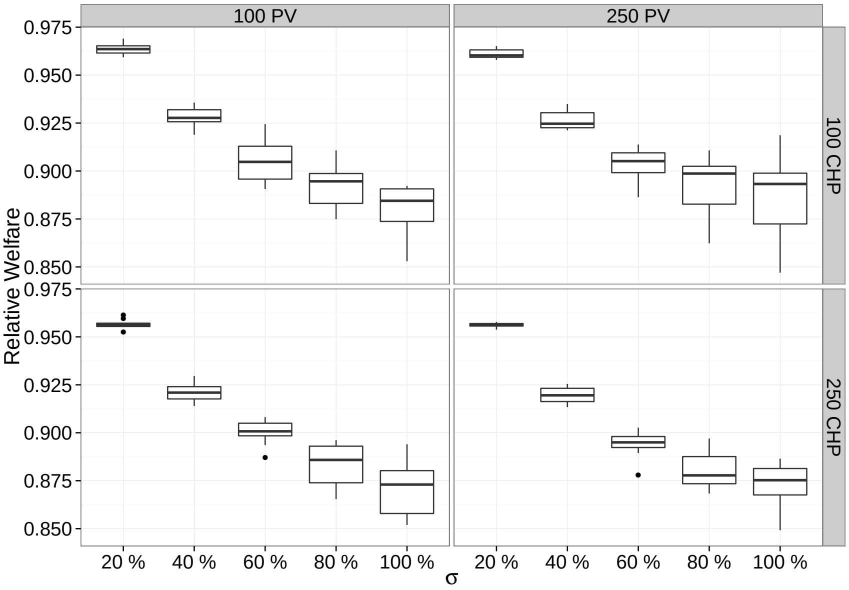


Figure 5.7: Relative welfare $\mathcal{W}'$ for the smart grid pricing scheme with saturation-level-independent utility and wavelet privacy

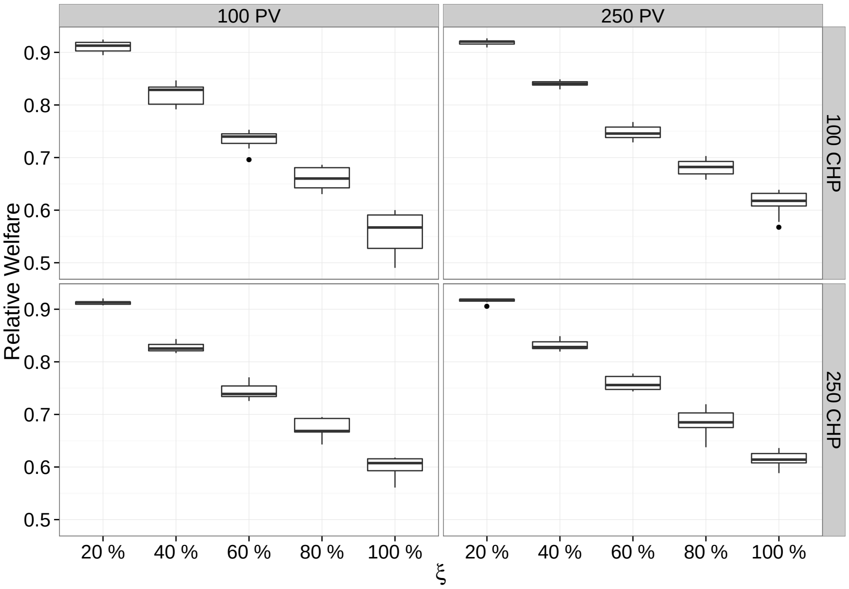


Figure 5.8: Relative welfare $\mathcal{W}'$ for traditional pricing scheme and IC wavelet privacy

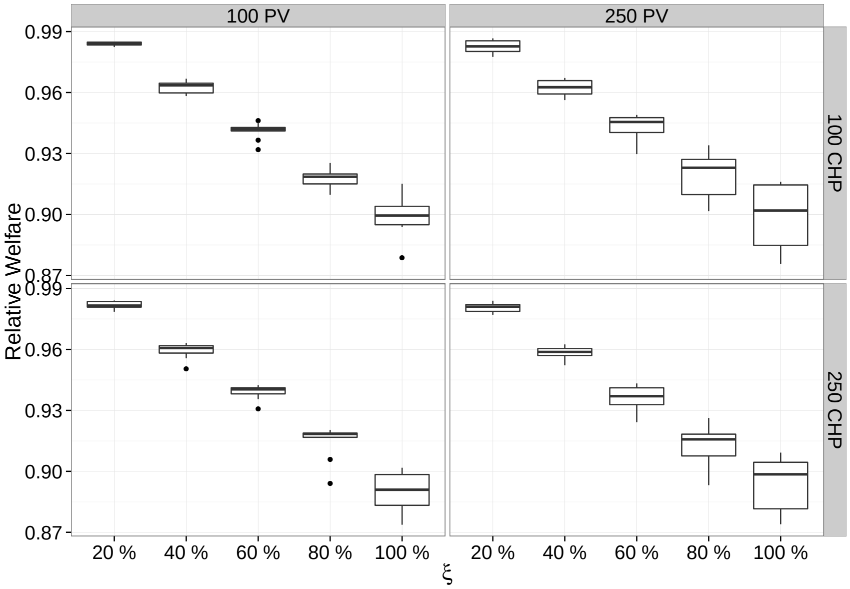


Figure 5.9: Relative welfare $\mathcal{W}'$ for the smart grid pricing scheme with saturation-level-dependent utility and IC wavelet privacy

increasing, whereas in the traditional scheme it remains constant. Inaccurate demand reports lead to additional emissions with traditional pricing, as the consumer limit price equals the balancing party price. In the smart grid pricing scheme, trades with the balancing party depend on the market situation. This explains the additional emissions in the traditional scheme. (Figures 5.2, 5.5, and 5.8 contain results for the traditional pricing scheme, Figures 5.4, 5.7, and 5.10 the results for the saturation-independent utility and Figures 5.3, 5.6, and 5.9 the results for the dependent utility.)

One important element of the smart grid is that consumers can express different valuations for electricity (see Chapter 1). Different valuations allow the expression of demand-side flexibility and facilitates distribution of renewable energy. The results exhibit that the expression of flexible valuations is also beneficial in the context of privacy enhancement.

Effect of Storage In theory, storage can help to reduce the welfare loss of privacy enhancement (see Lemma 8). This is because it is capable of storing bought electricity that exceeds the saturation level. Thus, it is not 'wasted' but can be used in times of undersupply. The results of the simulations quantify the actual impact on welfare in a real-world scenario. The results show that storage can reduce the privacy-induced welfare loss by 70% (Figs. 5.20–5.22). Next to this expected result, at least in qualitative terms, we make the following observations.

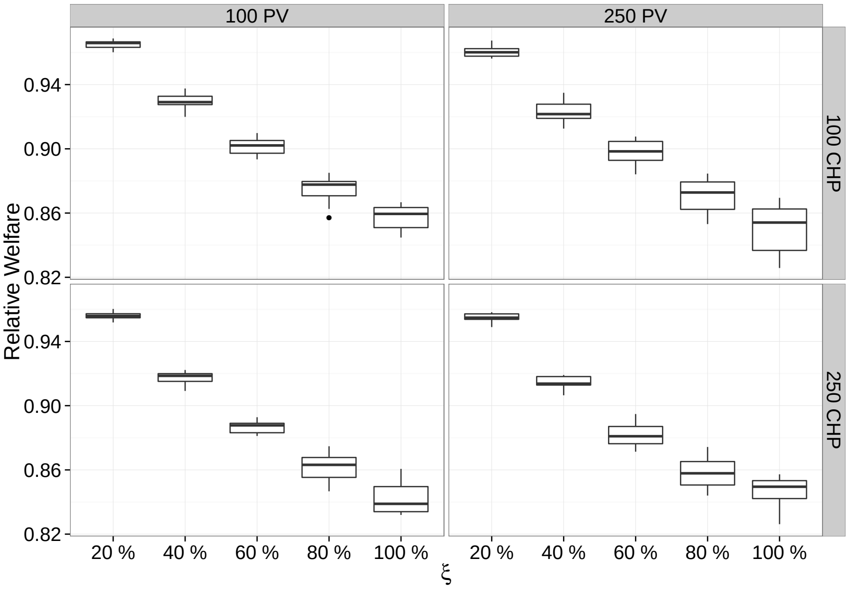


Figure 5.10: Relative welfare $\mathcal{W}'$ for the smart grid pricing scheme with saturation-level-independent utility and IC wavelet privacy

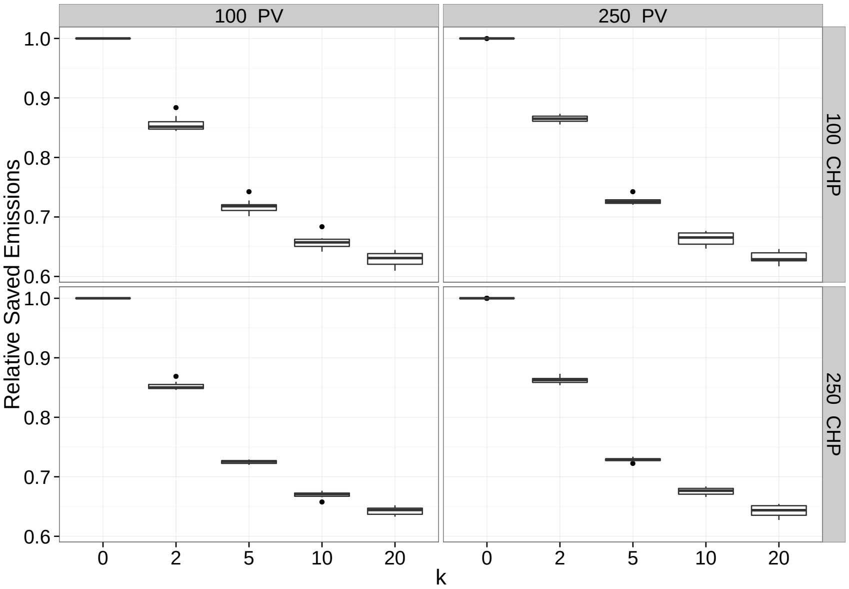


Figure 5.11: Relative saved CO_2 emissions Ω' for the traditional pricing scheme and k-anonymity

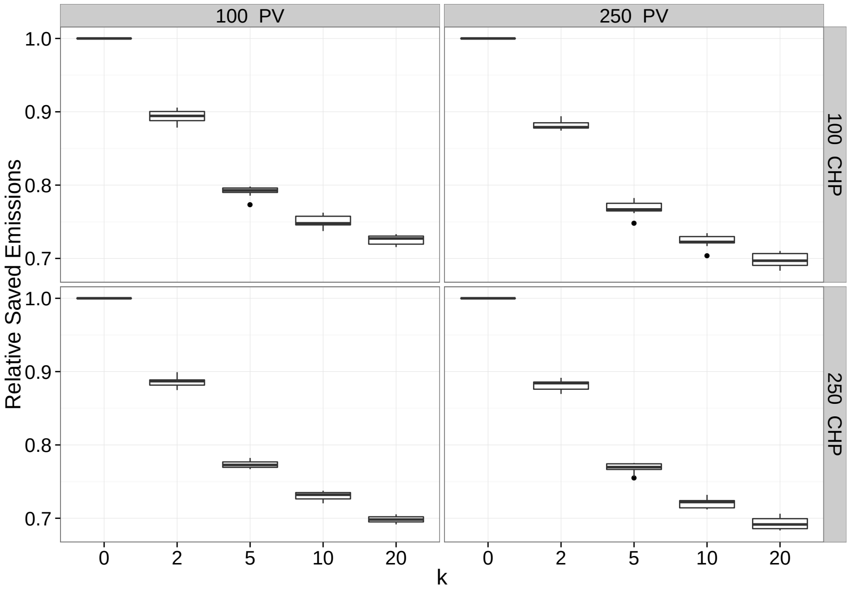


Figure 5.12: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-dependent utility and k-anonymity

R.23 *Small storage systems are sufficient to mitigate privacy costs.* Under our naïve privacy-driven storage operation strategy, the $2.5-kWh$ system is almost as efficient as the $5-kWh$ system. This suggests that privacy costs may be mitigated at comparably low costs.

R.24 *The value of storage is increasing in the privacy level.* Higher privacy level choices induce more frequent quantity mismatches, which are mitigated by the storage system.

R.25 *Privacy enhancement may increase welfare in the presence of storage.* With storage, relative welfare is not monotonically decreasing with the privacy level. This is because privacy enhancement can induce economic dispatching of the storage system: Electricity bought above the saturation level $\overline{x}_c(t)$ usually is relatively 'cheap' because of the concave utility function. In times of undersupply, the stored electricity is only used if less than $\overline{x}_c(t)$ is allocated on the market. Thus, the actual resulting utility of the stored electricity is much higher than the bid price in the former time slot.

R.26 *The value of storage is increasing in decentral generation capacity.* In the case of high decentral generation capacity, surplus energy stored will more often originate from these low-cost sources. Consequently, the use of the balancing party will decrease. This has a positive impact on social welfare.

These simulation results show that even small energy storage systems are very effective at mitigating the welfare loss resulting from privacy enhancement.

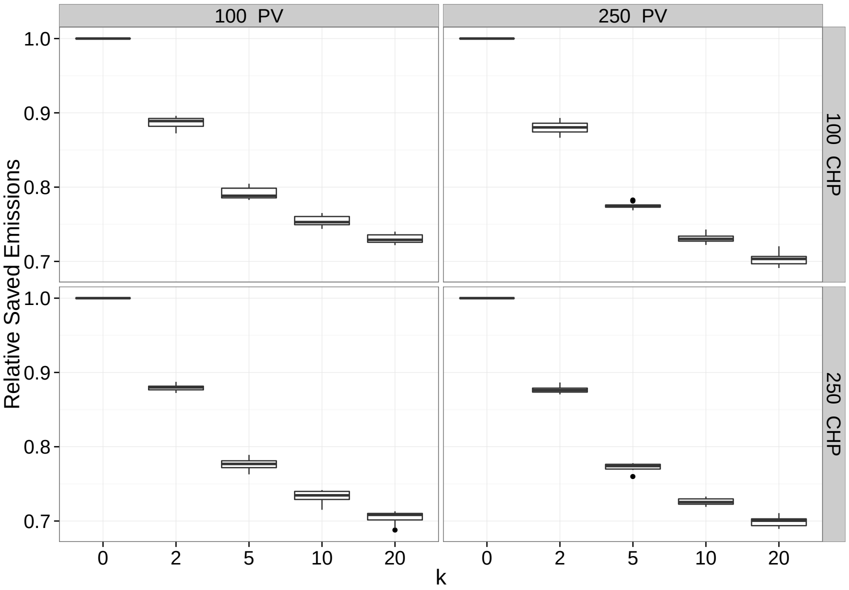


Figure 5.13: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-independent utility and k-anonymity

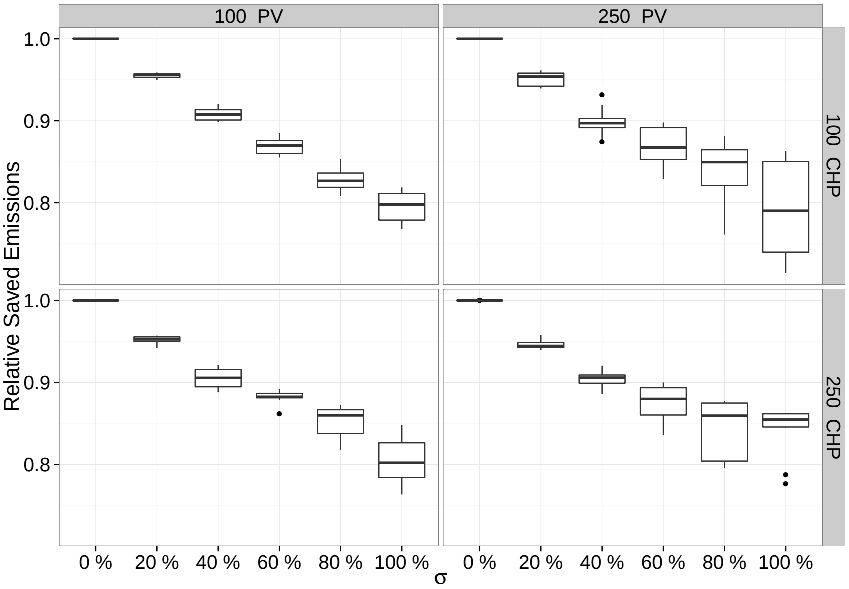


Figure 5.14: Relative saved CO_2 emissions Ω' for the traditional pricing scheme and wavelet privacy

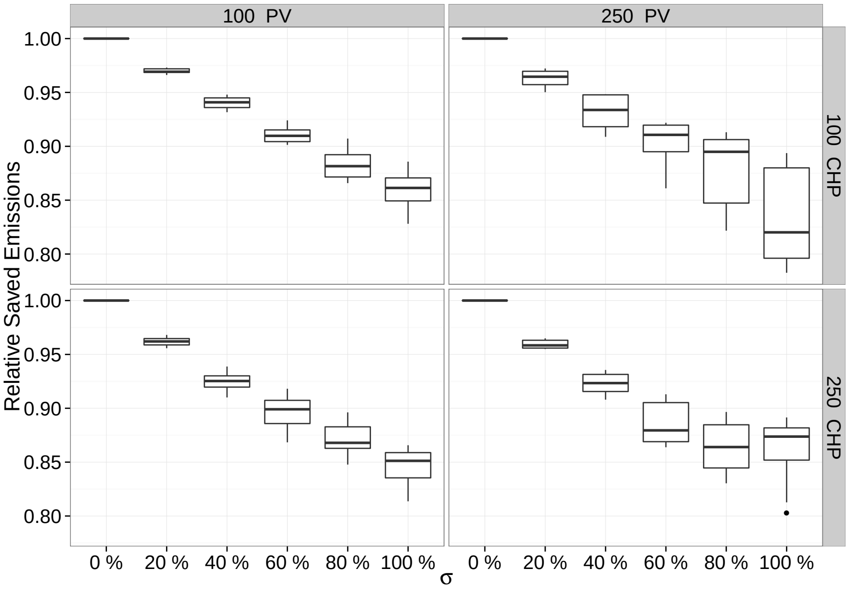


Figure 5.15: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-dependent utility and wavelet privacy

Impact of Demand-Side Flexibility We know from Lemma 9 that higher demand-side flexibility curbs the influence of the privacy-enhancement method on the welfare loss. For saturation-level-independent utility functions, the influence of demand-side flexibility is unpredictable (Lemma 9). In the traditional pricing scheme, we cannot model demand-side flexibility as every unit of energy is valuated equally. Thus, we only cover the dependent utility function.

R.27 *Demand-side flexibility can mitigate the welfare loss only to a limited extent.* The mitigating effect of demand-side flexibility is for the (IC) wavelet privacy-enhancing method lower than 3% (Fig. 5.23). Comparing that with the effects of storage, flexibility has only a little effect (Figs. 5.21 and 5.22). The higher relative welfare for low demand-side flexibility in one data point is the result of numerical instabilility (Fig. 5.23a). However, demand-side flexibility has a much larger effect for k-anonymity privacy enhancement. In particular, low demand-side flexibility results in severe welfare loss, whereas high flexibility has a lower positive effect. In comparison to storage (Fig. 5.20), mitigating effects are on a similar level.

Both demand flexibility and storage can mitigate the welfare loss of privacy enhancement. However, our results suggest that storage has a much larger potential. Load flexibility only helps to reduce the welfare loss if $\widetilde{x}_c(t) < \overline{x}_c(t)$. Storage in turn also helps in cases of over-allocation: When the privacy-enhancement method upward-adjusts the saturation level, the additional allocated electricity is not lost. If the saturation level is downward-adjusted, storage can help to mitigate the

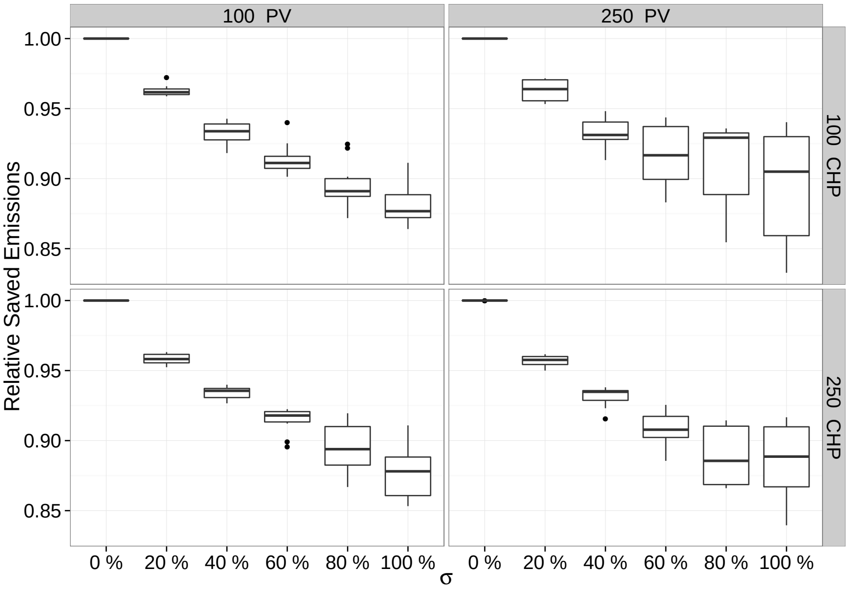


Figure 5.16: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-independent utility and wavelet privacy

welfare loss if $\mathcal{B}_c(t) > 0$.

Impact and Comparison of Privacy-Enhancement Methods As noted previously, the standard wavelet approach may induce strategic bidding on behalf of the consumers. This is because this privacy mechanism is not monotonously increasing. Here we want to analyze the variant of this algorithm for saturation-level-independent utility. The wavelet-privacy parameter σ varies between 0 and 100%. To ensure that the second privacy-enhancement method actually is monotonically increasing, we choose a fixed $\sigma = 30\%$ and vary the noise ξ only. We find that the welfare loss is more pronounced under our modified algorithm (Figs. 5.6, 5.7, 5.9, and 5.10).

R.28 *The IC wavelet-privacy method leads to a greater welfare loss.* In the standard wavelet privacy-enhancement method, σ influences both the choice of coefficients perturbed and the standard deviation of the noise. For the incentive compatible method, the perturbed coefficients are always the same. This is because σ is fixed to 30%. For a higher privacy level the wavelet privacy method perturbs fewer coefficients, resulting in a higher welfare. The additional welfare loss can be interpreted as the cost of establishing incentive compatibility. Although these costs remain negligible for privacy levels of up to 60%, they become more significant at higher privacy levels as the noise level is monotonically increasing.

R.29 *Increasing the privacy parameters of the (IC) wavelet privacy method have a linear effect*

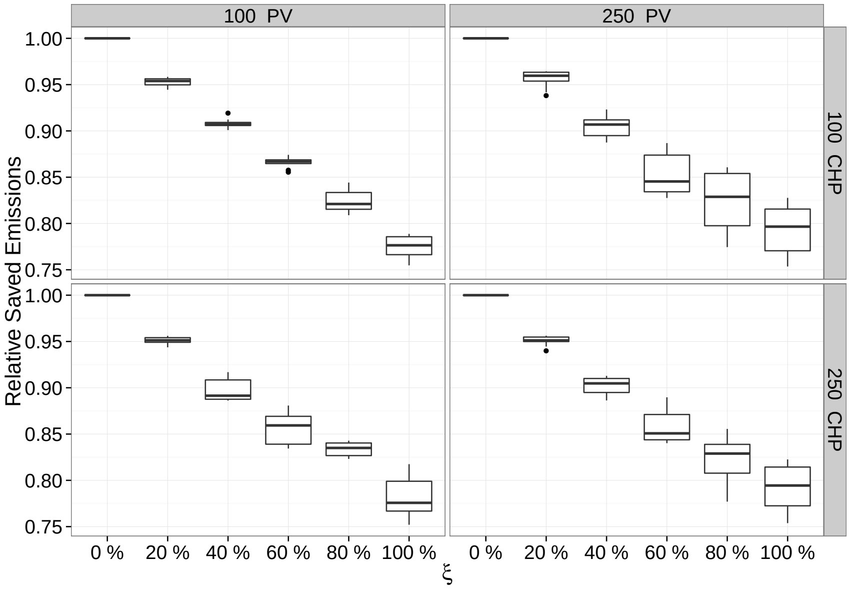


Figure 5.17: Relative saved CO_2 emissions Ω' for the traditional pricing scheme and IC wavelet privacy

on the relative welfare. The reason for the linear decrease is that each unit of electricity is valuated equally by consumers and by producers. Increasing σ or ξ leads to (in expectation) linearly increasing noise. Consequently, allocative efficiency decreases linearly and, because of the constant valuation of electricity units, also decreases linearly.

R.30 *The drop in relative welfare for k-anonymity flattens the higher k is.* Initially, finding groups of k similar time series requires changing the original values a lot. This especially holds for outliers. Severe modifications result in a large initial drop in welfare, i.e., compare $k = 2$ and $k = 5$ in Fig. 5.2. The investigated CER data set (Section 2.2.3) contains households from the same region, and most likely a large fraction has a similar behavior. The larger the chosen k groups, the more similar the whole data set becomes, in alignment with most households. This explains why the drop from $k = 5$ to $k = 10$ anonymity is lower than the one from $k = 2$ and $k = 5$.

In summary, different modification schemes of privacy-enhancement methods obviously have a different effect. In terms of privacy level, privacy methods are not necessarily comparable. In particular, choosing k-anonymity may be suitable for the privacy preferences of outstanding time series, e.g., households that consume a lot of electricity, because such time series are severely modified when putting them into a group. In turn, this kind of privacy may not be enough for households that have common consumption patterns. However, we can quantify the negative impact on data quality of the methods. Thus, it is worthwhile to compare the impacts of different

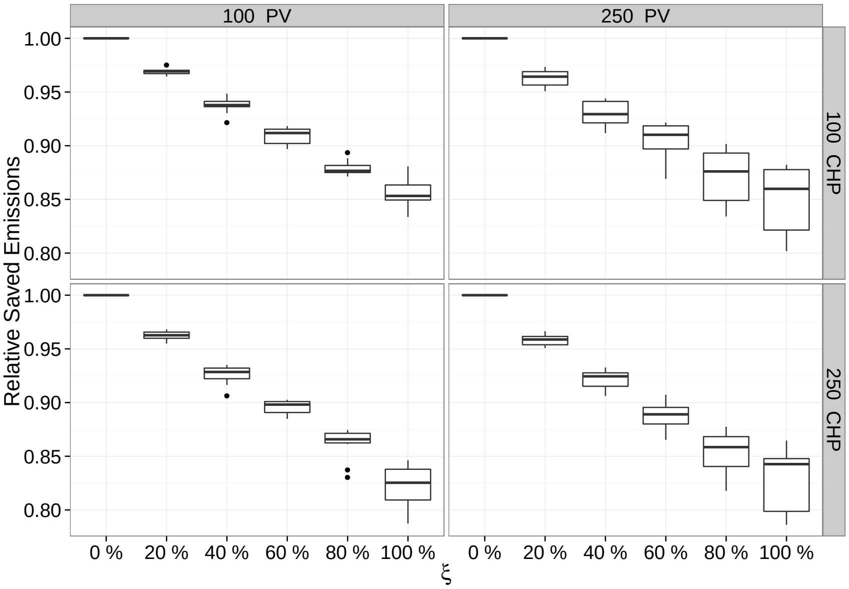


Figure 5.18: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-dependent utility and IC wavelet privacy

privacy-enhancement methods.

5.3.5 Numerical Results: PACTS

As previously discussed, applying provable privacy methods results in a complete loss of utility (see Section 3.1.2). Thus, in the following we are especially interested in whether PACTS can preserve the data quality necessary for a local energy market. Reconsider that hiding 'Light', 'Refrigerator' and 'Microwave' requires the same abstraction with increasing noise.

Effect of Market Structure We analyzed the effects on relative welfare (Figs. 5.24–5.26) and saved CO_2 emissions (Figs. 5.27–5.29) with different numbers of suppliers.

R.31 *PACTS, in combination with the traditional pricing scheme, results in a useless market.* For many secrets, PACTS leads to negative relative welfare, and for most configurations to relative welfare below 50%. Allocative inefficiency may lead to negative consumer surplus, as each unit of electricity is valued equally. In particular, each additionally bought unit reduces the surplus linearly because utility and limit prices are fixed (Fig. 5.24). Still, up to 75% of the emissions are saved in the presence of PACTS (Fig. 5.27). However, welfare values indicate that consumers will suffer a loss in such a market. Consequently, they will not participate. In this case, provable privacy guarantees lead to a useless data set for a local energy market.

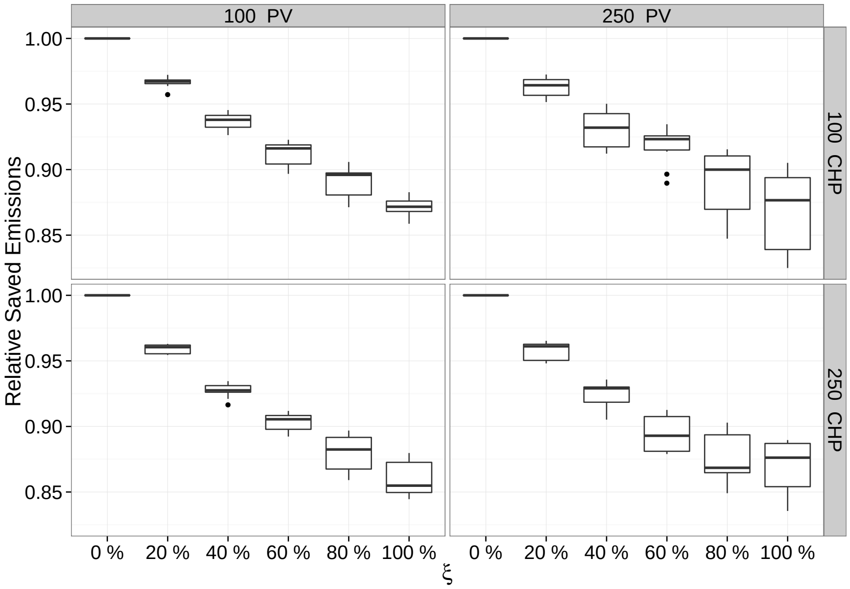


Figure 5.19: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-independent utility and IC wavelet privacy

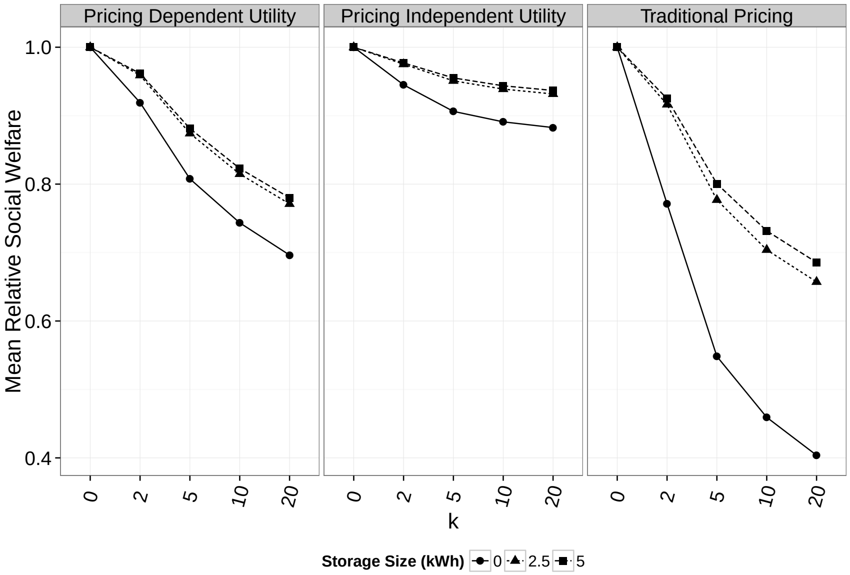


Figure 5.20: Relative welfare for k-anonymity with storage, 250 PV, and CHP sites present

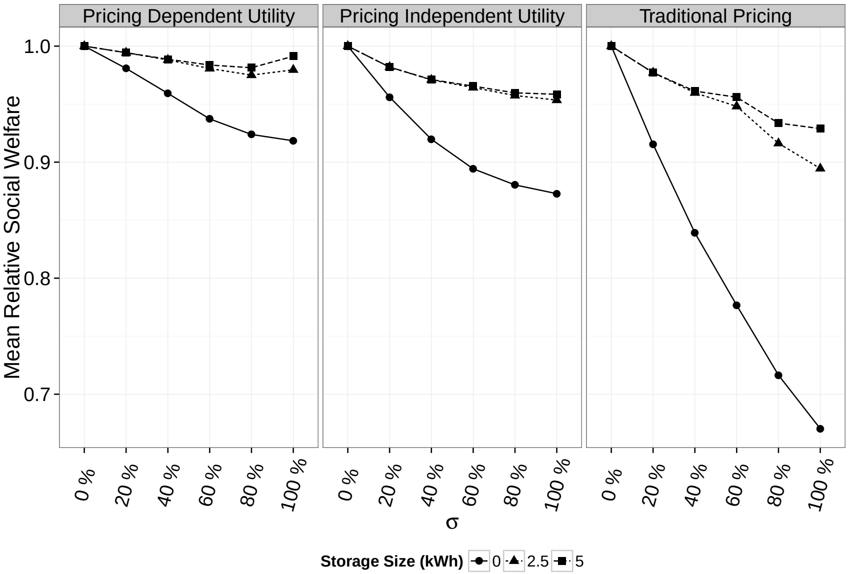


Figure 5.21: Relative welfare for wavelet privacy with storage, 250 PV, and CHP sites present

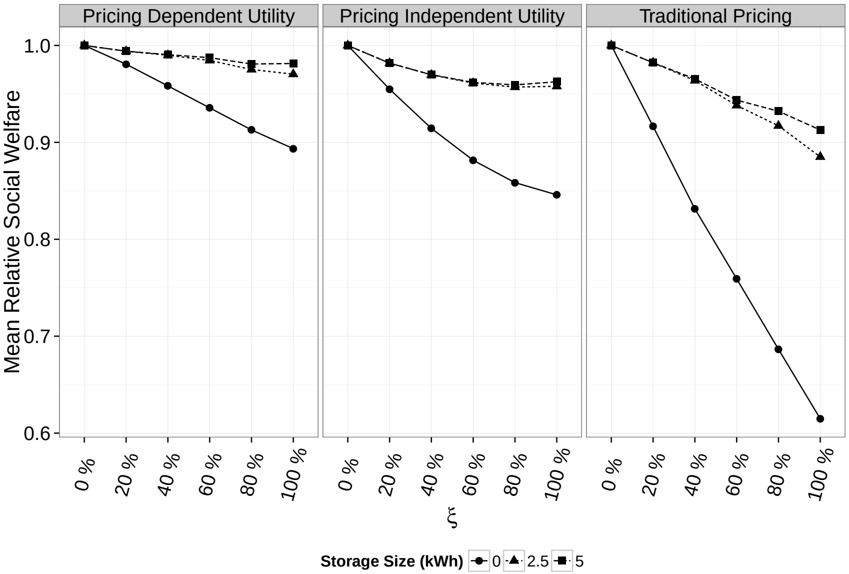


Figure 5.22: Relative welfare for IC wavelet privacy with storage, 250 PV, and CHP sites present

In the following, we will refer only to the smart grid pricing scheme, as PACTS is not applicable in a local energy market.

R.32 *PACTS is feasible in a local energy market with the smart grid pricing scheme.* Depending on the secret (with different noise levels), relative welfare differs, but in many cases it is above 60% and up to 85%. In the evaluated scenario, guaranteeing privacy leads to lower relative welfare compared to the other privacy-enhancement methods considered. Considering that we have assumed the 'worst case' scenario in that each secret is hidden in the complete time span, we can argue that the loss of welfare with PACTS is acceptable. The same conclusions also hold for the saved emissions (Figs. 5.25 and 5.26 contain the relative welfare, Figs. 5.28 and 5.29 contain the relative saved emissions).

R.33 *Effect on welfare and saved emissions of PACTS can depend on the market structure.* In particular, we consider the hiding the 'Wakeup Hour' respectively 'Bedtime Hour' feature in Figure 5.26. In the morning, when most households get up, as well as in the evening, most of the renewable electricity is produced by CHP sites, as the sun intensity is rather low. The more CHP sites, the more consumer surplus is achieved during that time. In turn, the more welfare is lost when hiding these specific secrets. The similar effect can be seen for the saved emissions (Fig. 5.29).

In contrast to the (IC) wavelet and k-anonymity privacy-enhancing approaches, in which uniform 'noise' is applied, PACTS may have different effects depending on the correlation between the secrets and number of suppliers. Estimating the loss of welfare becomes harder, but also allows the consumer to define less rigorous privacy requirements during peak times.

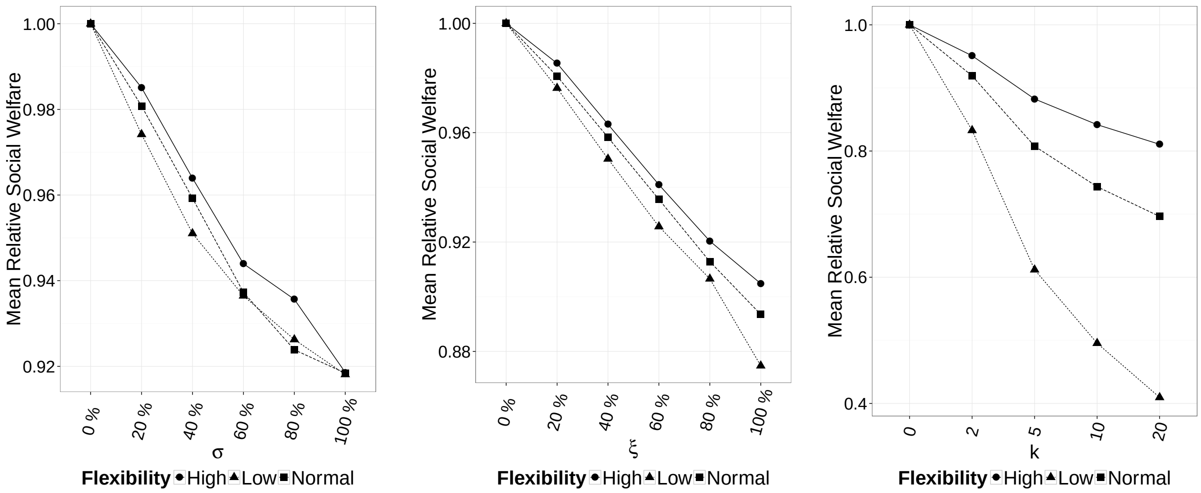
Effect of Storage We also evaluated the impact of storage on relative welfare with PACTS. Figure 5.30 contains those results.

R.34 *The effect of storage on relative welfare increases as the relative welfare without storage decreases.* For instance, the highest mitigating effect of storage is the case in which the state of the 'Refrigerator' is hidden. With the help of storage, the relative welfare is higher than 50% in all cases, and higher than 75% with the saturation-level-independent utility.

These results underline, that storage can have a significant effect on relative welfare when applying privacy-enhancement methods in local energy markets. Even in cases in which the hiding of secrets results in a severe welfare loss, storage mitigates a large fraction. The results clearly show that provable privacy guarantees can be achieved for consumers in a local energy market with acceptable loss of welfare.

Impact of Demand-Side Flexibility We also tested PACTS with different flexibility levels. Depending on the secret hidden, flexibility strongly influences the relative welfare.

R.35 *High demand-side flexibility can increase the relative social welfare up to* 10%*, and a low flexibility reduces welfare up to* 40% In general, PACTS adds more noise than other considered



(a) Wavelet

(b) IC Wavelet

(c) k-Anonymity

Figure 5.23: Impact of demand-side flexibility in the smart grid pricing scheme with saturation-level-dependent utility

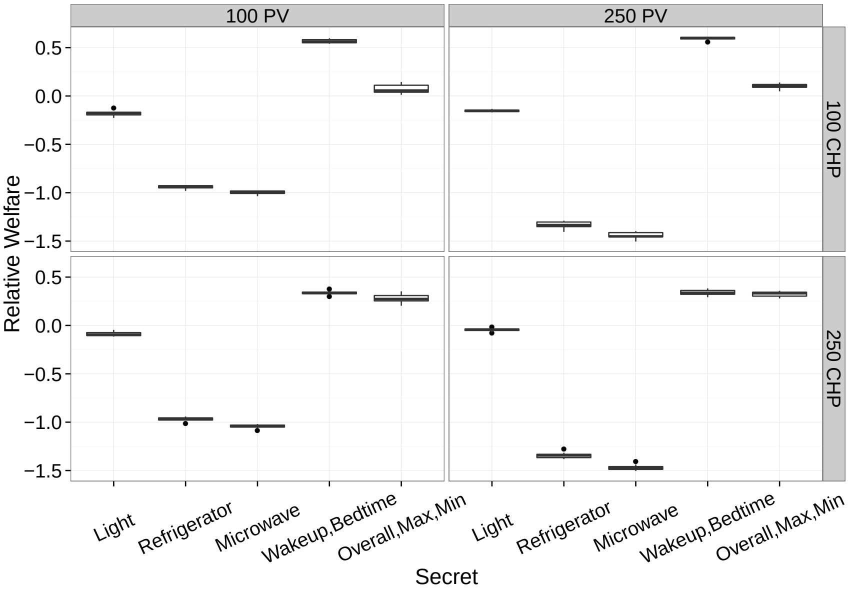


Figure 5.24: Relative welfare for the traditional pricing scheme and PACTS

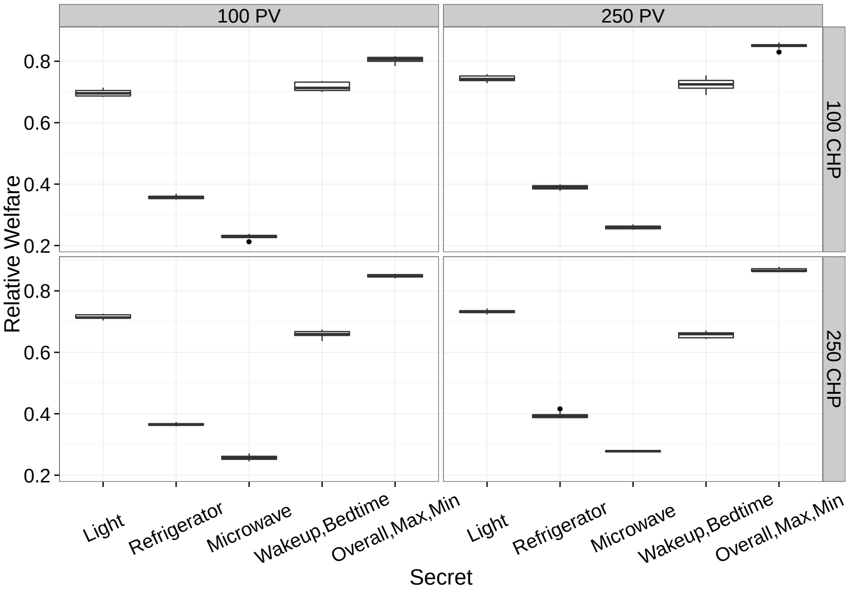


Figure 5.25: Relative welfare for the smart grid pricing scheme with saturation-level-dependent utility and PACTS

privacy-enhancing methods. Consequently, different flexibilities have a stronger effect on welfare. We find that, in particular, a low flexibility leads to a sharp drop in social welfare (Fig. 5.31).

Our results suggest that households with a low demand-side flexibility have a severe decrease in social welfare when applying PACTS. In contrast, even if the impact is smaller, a high demand-side flexibility preserves a significant fraction of welfare.

5.4 Conclusions

Privacy-aware local energy markets are a promising approach for matching renewable supply and demand of private households. However, the potential effects of privacy enhancement on the market outcome have so far remained vague. We provide a characterization of relevant privacy-enhancement properties when applied in a market scenario. Under certain assumptions, market mechanisms can retain incentive compatibility in the presence of privacy enhancement.

For approaches like k-anonymity and (IC) wavelet privacy, our numerical analyses show that loss of relative welfare and saved emissions are low. In combination with storage endowments, the loss of welfare is below 5%, in many cases even for strong privacy requirements. The saved emissions are still approximately 85% of the maximum amount of saved emissions.

As expected, provable privacy guarantees of PACTS result in a stronger decrease in relative

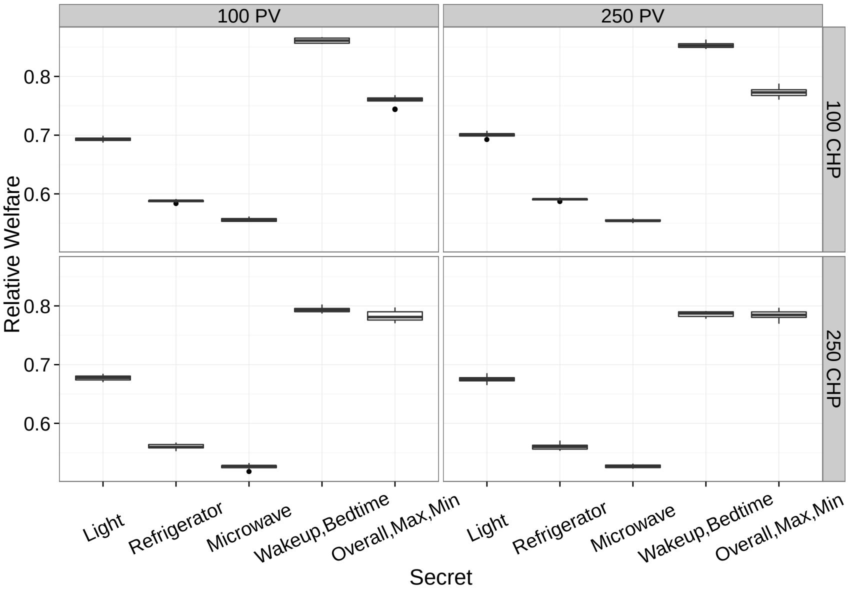


Figure 5.26: Relative welfare for the smart grid pricing scheme with saturation-level-independent utility and PACTS

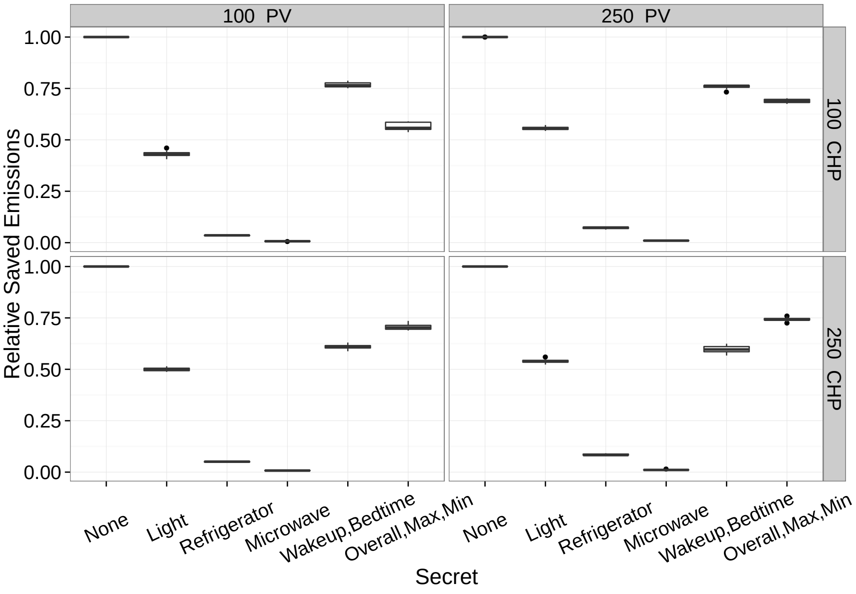


Figure 5.27: Relative saved CO_2 emissions Ω' for the traditional pricing scheme and PACTS

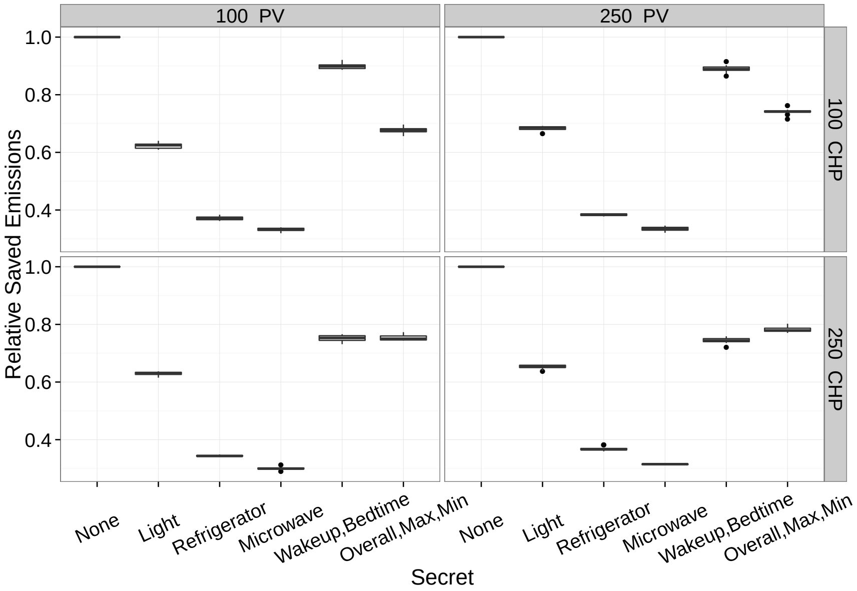


Figure 5.28: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-dependent utility and PACTS

welfare and saved emissions. However, when the hidden secrets are chosen carefully, a large fraction of welfare and emissions are retained. The presence of storage and a high demand-side flexibility improve the results. Thus, providing utility and provable privacy guarantees is possible in the local energy market scenario.

The overall conclusion is that privacy-enhancement methods are applicable in local energy markets, including private households. From an economic perspective, the negative allocative effects are low and controllable, whereas privacy enhancement significantly increases the privacy protection of participating individuals. From a computer science perspective, these markets are a meaningful performance indicator for the utility of privacy-enhancement methods.

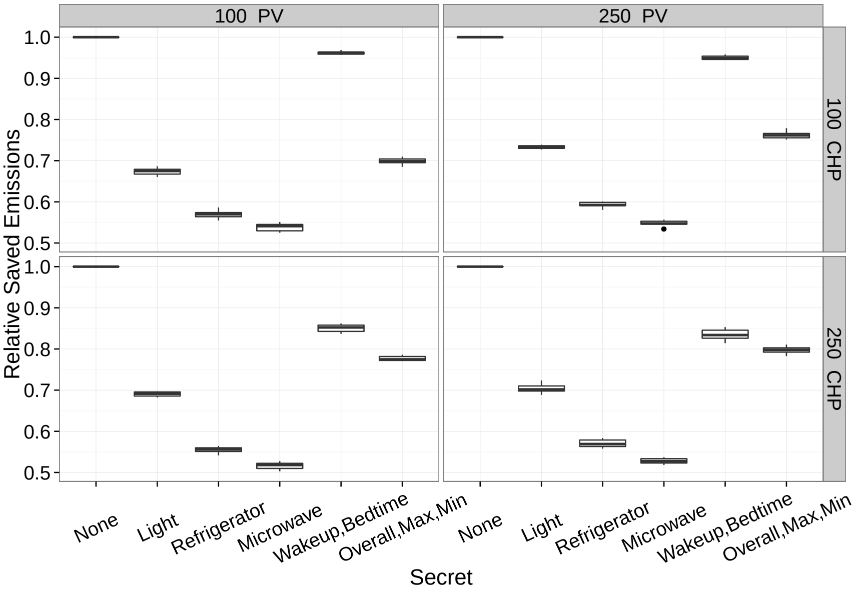


Figure 5.29: Relative saved CO_2 emissions Ω' for the smart grid pricing scheme with saturation-level-independent utility and PACTS

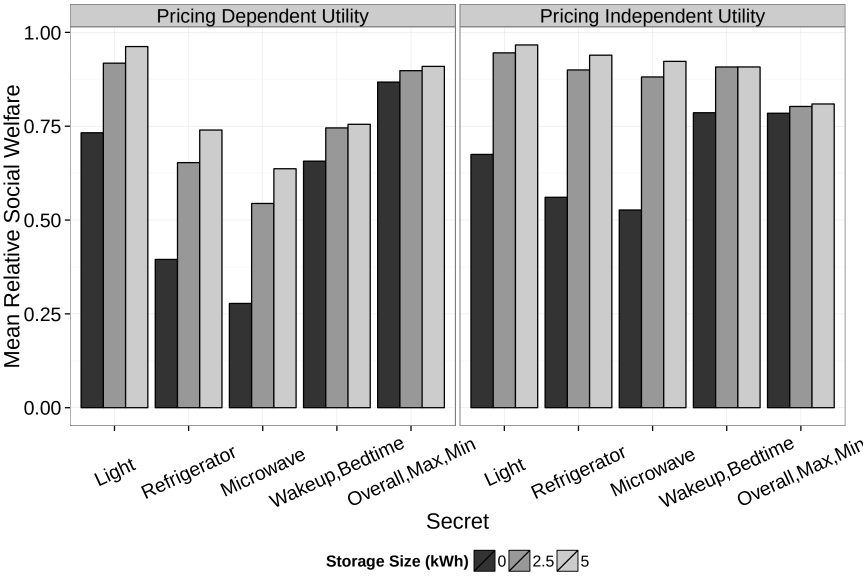


Figure 5.30: Relative welfare for PACTS with storage, 250 PV, and CHP sites present

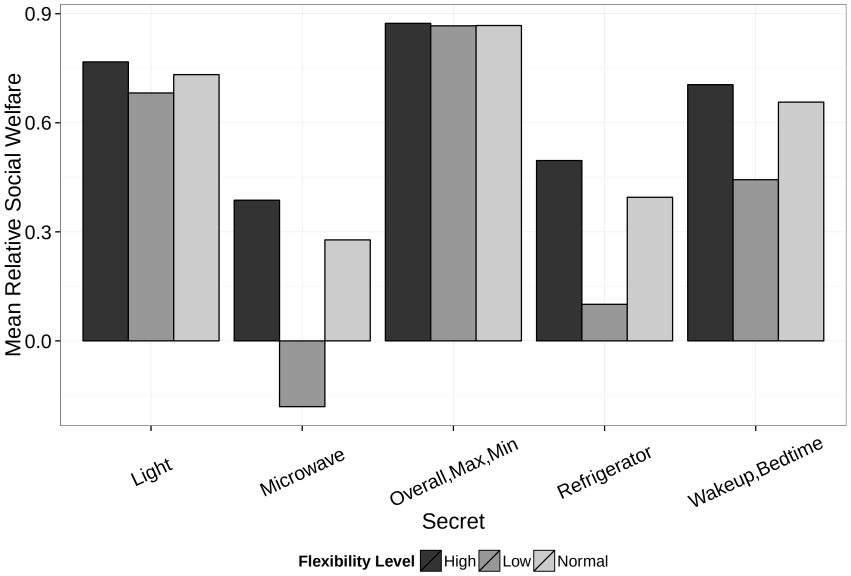


Figure 5.31: Relative welfare for PACTS with different flexibility levels, 250 PV, and CHP sites present

Chapter 6

Conclusion and Future Work

In many cases personal data are necessary for applications to be beneficial for society. One important example is the smart grid, in which access to fine-grain personal consumption data promises interesting new applications that reduce emissions and guarantee security of supply. However, arbitrary access to personal data puts privacy at risk. In this dissertation, we investigated the trade-off between privacy and utility, and strove to find a way of to fulfill both privacy and data requirements. The following chapter concludes the dissertation and summarizes the most important contributions and findings as well as an outlook on interesting future work.

6.1 Summary

In this dissertation, we investigated the privacy threats related to the time series of personal data. There exists a growing number of personal time-series data, e.g., GPS trajectories of mobile devices or energy consumption measured by smart meters in private households. As an application domain for the evaluation of our findings, we chose the smart grid scenario with the time series of smart meters. Legislation restricts access and requires special treatment for data that reflect living conditions and that can be assigned to individuals with minimal effort. Time series of personal data reflect living conditions, e.g., the wake-up and bedtime hours of households. Furthermore, as our first contribution, we have shown with a systematical method that smart-meter time series can be assigned to individuals with the help of features that describe consumption patterns with minimal effort. Consequently, the actual data (even without any personal identifier) is subject to privacy legislation and requires special treatment. This evaluation shows that a large fraction of households can be re-identified with minimal computational effort. Instead of striving for accurate re-identification, we have shown a system that provides a systematic way to do this, with minimal effort and with success rates that are significantly higher than ‘random guessing’. It is not surprising that re-identification rates differ, e.g., an increasing number of households considered reduce the rates because the likeliness of similar households increases. Additionally, features also have a different identifying degree, e.g., wake-up hours are more identifying for a shift worker in a data set, with households working on a nine-to-five basis. In addition to the system itself, we also contribute features with different identifying degrees.

As a second contribution, we proposed PACTS, a privacy-enhancement method that gives provable guarantees for individually defined privacy preferences. We have already seen that time series of smart-meter data contain a lot of different sensitive information. This includes features for re-identification and information about running devices. With PACTS, individuals decide independently of others, which information they deem sensitive. Informational self-determination requires strong guarantees regarding the removed information. Thus, PACTS gives provable privacy guarantees. PACTS provides ϵ-Pufferfish privacy guarantees for individually published time series of smart-meter data. Our evaluation has shown that PACTS is capable of removing a number of objectively chosen privacy requirements. Like other privacy-enhancement methods, PACTS modifies the data set. Although privacy is guaranteed with respect to the given assumptions, it is still questionable whether the modified data set still provides utility for applications.

The third contribution is the local energy market, an application-specific measure for the resulting data quality of privacy-enhancing methods. In such a market, private households as consumers and privately run renewable producers trade electricity in short time intervals. True demand reports of consumers allow the optimal allocative efficiency, maximizing the welfare and saved emissions. However, this puts the privacy at risk because the reported demands in short time slots match the time series of smart meters. Applying a privacy-enhancement method on the demand reports reduces the allocative efficiency. Comparing market outcomes with and without privacy enhancement allows one to measure how much the privacy-enhancement methods influence the data quality. In contrast to existing abstract measures like the L1-Norm, the local energy market results in intuitively understandable values, in particular CO_2 emissions and loss of welfare. The success of a local energy market depends on the typical effects for the smart grid, e.g., modifications during times of oversupply by a renewable source has a different effect than during times of undersupply. It is therefore possible to conclude that the local energy market, as a quality measure, is discriminative for smart grid applications in general.

Finally, we investigated the actual impact of privacy-enhancement methods on such markets. We found on a theoretical basis that market mechanisms keep important properties like incentive compatibility in the presence of privacy-enhancement methods. As numerical results, we investigated the actual impact of common related privacy-enhancement methods and PACTS. In addition, we have also seen that storage systems can significantly mitigate the negative effects of privacy enhancement. A consumer may invest in such a system to preserve privacy while still keeping high surplus. The overall conclusion is that privacy enhancement is applicable in markets such as the proposed local energy market. Effects are controllable and low in comparison to the achieved benefit for society.

This dissertation shows that it is possible to combine the contradictory goals of privacy and utility gained from access to personal data. We have shown that both theoretically and numerically in the smart grid scenario. Privacy-enhancement methods for time series give way for numerous applications that are beneficial to society and that require personal data. Individuals keep their right to informational self-determination when they define their privacy preferences.

6.2 Future Work

In the following, we elaborate on possible fields of future research, to continue the contributions in this dissertation.

6.2.1 Simplification of Privacy Requirements in PACTS

In Chapter 3 we introduced PACTS as a provable privacy approach that respects individual requirements. Users have to define discriminative pairs such as s_{pair} = (‘The flow heater is starting.’, ‘The flow heater is not starting.’). PACTS requires a transformation mechanism for time series in an abstracted representation and a distinct mapping for the abstract coefficients correlation to s_{pair}. An individual with very little technical knowledge would not be able to provide such transformations and definitions for arbitrary secrets. Therefore, to allow every individual the use of PACTS, we require an easier way to define such secrets and discriminative pairs. One might research the following possibilities:

- *Automatic learning:* With the help of locally installed smart meters, individuals can automatically learn signatures by manually and controlled use of devices that should be hidden. With the help of the learned signatures, we can determine a wavelet basis with methods such as lifting [3].
- *Central signature database:* Activities involving the same type of device usually lead to typical electricity consuming signatures. Although this also allows the extraction of information in smart-meter data, we could use this to build a database for ‘commonly secrets’. Such predefined database entries contain an understandable description as well as the required transformation mechanism and a mapping to coefficients in question. Individuals choose the activities they wish to hide, and include them in their privacy enhancement.

6.2.2 Utility Guaranty

In general, privacy enhancing methods focus on requirements from individuals regarding their privacy. The goal is to provide applications with meaningful data while still keeping the right of individual self-determination. On the other hand, applications also have requirements for the utility of the data, i.e., depending on the actual modifications data might not be usable at all. Thus, the requirement of such an application to be able to use the data would be, that it is only modified to a certain extent. For instance, one requirement can be that the moving average of all the time series in a data set is preserved during privacy-enhancement. If such requirements for data is not fulfilled, it leads to the following contradiction: A user publishes privacy-enhanced data for the benefit of society, but the data is useless for certain applications. Consequently, the benefit for society strives to zero. In this case, it is preferable not to publish the data at all, because it does not provide utility and this obviously minimizes the risk of any privacy breach.

One possibility to overcome the explained contradiction is to combine privacy with application requirements. If both cannot be matched, no data is published at all. Challenges for such a system are, amongst others, to specify a way to express application specific requirements and on the other hand, to make sure that both, privacy as well as utility requirements, are still respected.

6.2.3 Optimal Privacy and Welfare

With the help of the local energy market, we were able to quantify the welfare loss of different privacy-enhancement methods (see Chapters 4 and 5). We could predict the welfare loss for a given privacy-enhancing method and parameters. However, we were not able to define an 'optimal' way of privacy enhancement for an individual because we could not quantify the valuation for privacy. If we could specify such valuations, our market would be able to determine the optimal methods and parameters. Finding such valuations is challenging, as they depend on individual perception of privacy. However, valuations for privacy would shed light on the dependency between privacy and utility from an individual perspective. One possible way of valuating privacy, is explained in the following section.

Valuations for Privacy based on Identifying Degree

One promising approach to introduce a objective valuation is to rate the value of certain information by their identifying degree. We have seen in Chapter 2 that time series of power consumption values are identifying with the help of different features. Each feature represents a certain information of the individual household. We also found out, that all features differ in their identifying degree. On model for valuation would be: The higher the identifying degree the higher the valuation of an individual is to keep exactly this information private. That model depends on the actual data set, i.e., in a set consisting of time series from employees working at regular business hours the 'Average Wakeup Hour' is not as identifying as in a data set of shift workers. Thus, this model also incorporates perception with respect to the peer group of a household. This model still requires a reference point, since currently it only provides an order of valuations.

6.2.4 Extending the Application-Specific Data Quality Measure

In Chapter 4 we explained the benefits of an application specific measure such as the proposed local energy market. Applications in the same domain rely on similar characteristics of the data. Thus, we expect that the results of the local energy market are also discriminative for other smart grid applications. However, one might still be interested in actual values regarding a specific application rather than qualitative results derived from the impact on electronic markets. Investigating other applications if they are suitable as a measure for data quality will lead to a deeper understanding how privacy enhancement methods interact with actual applications. It is a promising line of research to analyze the following applications:

- Detailed usage data should optimize the planning of electricity network expansion. However, it is unknown how privacy enhancing methods influence possible results.
- Self-tracking devices measuring body functions such as heart rate as time series increase in popularity. Exchange and comparison of such data allow conclusions regarding health. It is also unknown, how privacy enhancement methods applied to self-tracking data changes results.

The proposed techniques of privacy enhancement can be used for such time series with none or minimal modifications. The major challenge of finding suitable measures for this applications remains.

Bibliography

[1] O. Abul, F. Bonchi, and M. Nanni. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. *IEEE 24th International Conference on Data Engineering*, pages 376–385, 2008.

[2] O. Abul, F. Bonchi, and M. Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910, 2010.

[3] T. Acharya and C. Chakrabarti. A Survey on Lifting-based Discrete Wavelet Transform Architectures. *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 42(3):321–339, 2006.

[4] T. Ackermann, G. Andersson, and L. Söder. Distributed generation: a definition. *Electric Power Systems Research*, 57(3):195–204, 2001.

[5] G. Acs and C. Castelluccia. I have a DREAM! (DIffeRentially PrivatE smart Metering). *Information Hiding*, pages 118–132, 2011.

[6] G. Aggarwal, T. Feder, and K. Kenthapadi. Anonymizing tables. *Lecture Notes in Computer Science*, pages 246–258, 2005.

[7] M. Barbaro, T. Zeller, and S. Hansell. A face is exposed for AOL searcher no. 4417749. *New York Times*, 2006.

[8] N. Batra, H. Dutta, and A. Singh. INDiC: Improved Non-intrusive Load Monitoring Using Load Division and Calibration. *12th International Conference on Machine Learning and Applications*, pages 79–84, 2013.

[9] E. Bitar and S. Low. Deadline differentiated pricing of deferrable electric power service. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4991–4997, 2012.

[10] F. Brandt. Fully private auctions in a constant number of rounds. *Computer Aided Verification*, pages 223–238, 2003.

[11] E. Buchmann, K. Böhm, T. Burghardt, and S. Kessler. Re-identification of Smart Meter data. *Personal and Ubiquitous Computing*, 17(4):653–662, 2012.

[12] E. Buchmann, S. Kessler, P. Jochem, and K. Böhm. The Costs of Privacy in Local Energy Markets. In *IEEE Conference on Business Informatics*, 2013.

[13] Bundesnetzagentur. Wettbewerbliche Entwicklungen und Handlungsoptionen im Bereich Zähl- und Messwesen und bei variablen Tarifen. 2010.

[14] Bundesrepublik Deutschland. Gesetz für den Vorrang Erneuerbarer Energien. *Bundesgesetzblatt*, I/2008:1754, 2008.

[15] Bundesrepublik Deutschland. Bundesdatenschutzgesetz (BDSG). pages 1–38, 2009.

[16] Bundesrepublik Deutschland. Kraft-Wärme-Kopplungsgesetz. *Bundesgesetzblatt*, I/2012: 1494, 2012.

[17] D. Bundestag. Gesetz über die Elektrizitäts-und Gasversorgung (Energiewirtschaftsgesetz-EnWG). *EnWG*, pages 1–94, 2005.

[18] M. Burrows and D. Wheeler. A Block-sorting Lossless Data Compression Algorithm. 1994.

[19] D. M. Burton. *The History of Mathematics: An Introduction.* 6 edition, 2007.

[20] R. Chen, B. C. Fung, N. Mohammed, B. C. Desai, and K. Wang. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, 2011.

[21] R. Coifman and M. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[22] T. Dalenius. Finding a needle in a haystack-or identifying anonymous census record. *Journal of official statistics*, 1986.

[23] A. Das. *Signal conditioning : an introduction to continuous wave communication and signal processing.* 2012.

[24] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics 41.7*, 41(7):909–996, 1988.

[25] M. De Paepe, P. D'Herdt, and D. Mertens. Micro-CHP systems for residential applications. *Energy Conversion and Management*, 47(18-19):3435–3446, 2006.

[26] Deutscher Bundestag. Verordnung über Rahmenbedingungen für den Messstellenbetrieb und die Messung im Bereich der leitungsgebundenen Elektrizitäts- und Gasversorgung (Messzugangsverordnung - MessZV). pages 1–7, 2008.

[27] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical society*, (April 2013):37–41, 1995.

[28] D. Donoho and P. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[29] C. Dwork. Differential privacy. *Automata, languages and programming*, pages 1–12, 2006.

[30] European Comission. The European Strategic Energy Technology Plan (SET Plan). URL `http://ec.europa.eu/energy/technology/set_plan/doc/setplan_brochure.pdf`.

[31] European Comission. European SmartGrids technology platform: vision and strategy for europe's electricity networks of the future. *Directorate for Research*, 2006.

[32] European Comission. SmartGrids SRA 2035 Strategic Research Agenda: Update of the SmartGrids SRA 2007 for the needs by the year 2035. *European Technology Platform SmartGrids*, (March), 2012. URL `http://www.smartgrids.eu/documents/sra2035.pdf`.

[33] European Parliament. Directive 95/46/EC. *Official Journal of the European Communities*, L 281/31, 1995.

[34] European Parliament. Directive 96/92/EC of the European Parliament and of the Council of 19 December 1996 concerning common rules for the internal market in electricity. *Official Journal No. L*, 27, 1997.

[35] Eurostat. Household electricity prices in the EU27 rose by 6.6% and gas prices by 10.3%, 2013. URL `http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/8-27052013-AP/EN/8-27052013-AP-EN.PDF`.

[36] N. Farag and M. S. Krishnan. The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency. 30(1):13–28, 2014.

[37] H. Farhangi. The path of the smart grid. *IEEE Power and Energy Magazine*, 8(1):18–28, 2010.

[38] A. Faruqui, D. Harris, and R. Hledik. Unlocking the 53 billion Euro savings from smart

meters in the EU: How increasing the adoption of dynamic tariffs could make or break the EU's smart grid investment. *Energy Policy*, 38(10):6222–6231, 2010.

[39] S. Finster and I. Baumgart. Pseudonymous Smart Metering without a Trusted Third Party. In *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pages 1723–1728, 2013.

[40] S. Finster and I. Baumgart. SMART-ER: Peer-based privacy for smart metering. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 652–657, 2014.

[41] H. Gharavi and R. Ghafurian. Smart Grid: The Electric Energy System of the Future [Scanning the Issue]. *Proceedings of the IEEE*, 99(6):917–921, 2011.

[42] H. Goncalves and A. Ocneanu. Unsupervised Disaggregation of Appliances using aggregated Consumption Data. In *The 1st KDD Workshop on Data Mining Applications in Sustainability*, 2011.

[43] U. Greveler, B. Justus, and D. Loehr. Multimedia content identification through smart meter power usage profiles. *5th International Conference on Computers, Privacy and Data Protection*, 2012.

[44] M. Gruteser and D. Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services - MobiSys '03*, pages 31–42, 2003.

[45] G. Hart. Nonintrusive Appliance Load Monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[46] R. Hledik. How Green Is the Smart Grid? *The Electricity Journal*, 22(3):29–41, 2009.

[47] K.-l. Hui, H. H. Teo, and S.-y. T. Lee. The Value of Privacy Assurance : An Exploratory. *MIS Quarterly*, 31(1):19–33, 2014.

[48] F. Hvelplund. Renewable energy and the need for local energy markets. *Energy*, 31(13): 2293–2302, 2006.

[49] D. Ilic, P. G. Da Silva, S. Karnouskos, and M. Griesemer. An energy market for trading electricity in smart grid neighbourhoods. *6th Internation Conference on Digital Ecosystems Technology*, pages 1–6, 2012.

[50] A. Ipakchi and F. Albuyeh. Grid of the future. *IEEE Power and Energy Magazine*, 7(2): 52–62, 2009.

[51] Irish Social Science Data Archive. Electricity Customer Behaviour Trial, 2012. URL `http://www.ucd.ie/issda/`.

[52] A. Jäger-Waldau. Photovoltaics and renewable energies in Europe. *Renewable and Sustainable Energy Reviews*, 11(7):1414–1437, 2007.

[53] M. Jawurek, M. Johns, and K. Rieck. Smart metering de-pseudonymization. In *Proceedings of the 27th Annual Computer Security Applications Conference on - ACSAC '11*, page 227, 2011.

[54] G. Kalogridis, C. Efthymiou, S. Denic, T. Lewis, and R. Cepeda. Privacy for smart meters: Towards undetectable appliance load signatures. In *First IEEE International Conference on*

Smart Grid Communications, pages 232–237, 2010.

[55] R. H. Katz, D. E. Culler, S. Sanders, S. Alspaugh, Y. Chen, S. Dawson-Haggerty, P. Dutta, M. He, X. Jiang, L. Keys, A. Krioukov, K. Lutz, J. Ortiz, P. Mohan, E. Reutzel, J. Taneja, J. Hsu, and S. Shankar. An information-centric energy infrastructure: The Berkeley view. *Sustainable Computing: Informatics and Systems*, 1(1):7–22, 2011.

[56] S. Kessler, E. Buchmann, T. Burghardt, and K. Böhm. Pattern-sensitive Time-series Anonymization and its Application to Energy-Consumption Data. *Open Journal of Information Systems (OJIS)*, 1(1):3–22, 2014.

[57] S. Kessler, C. Flath, and K. Böhm. Allocative and strategic effects of privacy enhancement in smart grids. *Information Systems*, 2014.

[58] S. Kessler, E. Buchmann, and K. Böhm. Deploying and Evaluating Pufferfish Privacy for Smart Meter Data (Technical Report). *Karlsruhe Reports in Informatics*, 1, 2015.

[59] D. Kifer and A. Machanavajjhala. No free Lunch in Data Privacy. *Proceedings of the International Conference on Management of Data*, page 193, 2011.

[60] D. Kifer and A. Machanavajjhala. A Rigorous and Customizable Framework for Privacy. *31st Symposium on Principles of Database Systems*, page 77, 2012.

[61] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han. Unsupervised Disaggregation of low frequency Power Measurements. Number i. HP Labs Tech. Report, 2010.

[62] J. Kolter and M. Johnson. REDD: A public data set for energy disaggregation research. *Workshop on Data Mining Applications in Sustainability*, (1):1–6, 2011.

[63] Konferenz der Datenschutzbeauftragten des Bundes und der Länder. Orientierungshilfe datenschutzgerechtes Smart Metering Juni 2012, Konferenz der Datenschutzbeauftragten, 2012.

[64] C. Kost, T. Schlegl, J. Thomson, S. Nold, and J. Mayer. Studie Stromgestehungskosten Erneuerbare Energien, 2012.

[65] S. Lamparter, S. Becher, and J. Fischer. An agent-based market platform for Smart Grids. *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 1689–1696, 2010.

[66] Landis + Gyr. Landis + Gyr E450 (Data Sheet). . URL `http://www.landisgyr.be/webfoo/wp-content/uploads/2012/09/D000028191_E450_f_en.pdf`.

[67] Landis + Gyr. Electricity Meters Ferraris, . URL `http://www.landisgyr.com/webfoo/wp-content/uploads/product-files/D000011432_Ferraris_d_en.pdf`.

[68] C. Laughman, K. Lee, and R. Cox. Power signature analysis. *Power and Energy Magazine, IEEE*, (april 2003), 2003.

[69] T. Li. t-closeness: Privacy beyond k-anonymity and ldiversity. *International Conference on Data Engineering (ICDE)*, (2), 2007.

[70] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 517, 2009.

[71] M. Liserre, T. Sauter, and J. Hung. Future Energy Systems: Integrating Renewable En-

ergy Sources into the Smart Power Grid Through Industrial Electronics. *IEEE Industrial Electronics Magazine*, 4(1):18–37, 2010.

[72] G. Locke and P. Gallagher. NIST framework and roadmap for smart grid interoperability standards, release 1.0. *National Institute of Standards and Technology*, 2010.

[73] S. L. Lyon. Privacy Challenges Could Stall Smart Grid. *Matter Network*, 2010.

[74] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L -diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[75] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets Practice on the Map. *IEEE 24th International Conference on Data Engineering*, 00:277–286, 2008.

[76] A. Madhavan. Security Prices and Market Transparency. *Journal of Financial Intermediation*, 5(3):255–283, 1996.

[77] S. Massoud Amin and B. Wollenberg. Toward a smart grid: power delivery for the 21st century. *Power and Energy Magazine, IEEE*, 3(5):34–41, 2005.

[78] P. McDaniel and S. McLaughlin. Security and Privacy Challenges in the Smart Grid. *IEEE Security & Privacy Magazine*, 7(3):75–77, 2009.

[79] E. McKenna, I. Richardson, and M. Thomson. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, 41:807–814, 2011.

[80] K. Mets, T. Verschueren, W. Haerick, C. Develder, and F. De Turck. Optimizing smart energy control strategies for plug-in hybrid electric vehicle charging. *2010 IEEE/IFIP Network Operations and Management Symposium Workshops*, pages 293–299, 2010.

[81] A. Mohsenian-Rad. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE Transactions on Smart Grid*, 1(3): 320–331, 2010.

[82] M. Mokbel, C. Chow, and W. Aref. The new Casper: query processing for location services without compromising privacy. In *Proceedings of the 32nd international conference on Very large data bases*, number 1, pages 763–774, 2006.

[83] A. Molina-Markham and P. Shenoy. Private Memoirs of a Smart Meter. *Proceedings of the BuildSys*, pages 61–66, 2010.

[84] K. Moslehi and R. Kumar. A Reliability Perspective of the Smart Grid. *IEEE Transactions on Smart Grid*, 1(1):57–64, 2010.

[85] C. Nabe, C. Beyer, and N. Brodersen. Ökonomische und technische Aspekte eines flächendeckenden Rollouts intelligenter Zähler. *Ecofys im Auftrag der Bundesagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen*, 2009.

[86] M. Nergiz and M. Atzori. Towards Trajectory Anonymization: a Generalization-based Approach. *SIGSPATIAL ACM GIS*, 2(106):47–75, 2008.

[87] M. Pagano and A. Roell. Transparency and Liquidity: A Comparison of Auction and Dealer Markets with Informed Trading. *The Journal of Finance*, 51(2):579, 1996.

[88] S. Papadimitriou, F. Li, and G. Kollios. Time series compressibility and privacy. *33rd Conference on Very Large Databases*, pages 459–470, 2007.

[89] D. C. Parkes, M. O. Rabin, S. M. Shieber, and C. Thorpe. Practical secrecy-preserving, verifiably correct and trustworthy auctions. *Electronic Commerce Research and Applications*, 7(3):294–312, 2008.

[90] S. Parsons, J. a. Rodriguez-Aguilar, and M. Klein. Auctions and bidding. *ACM Computing Surveys*, 43(2):1–59, 2011.

[91] J. M. Pearce. Photovoltaics — a path to sustainable futures. *Futures*, 34(7):663–674, 2002.

[92] D. B. Percival and A. Walden. *Wavelet Methods for Time Series Analysis.* 2006.

[93] L. Pérez-Lombard, J. Ortiz, and C. Pout. A review on buildings energy consumption information. *Energy and Buildings*, 40(3):394–398, 2008.

[94] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-divanis. Select-Organize-Anonymize : A framework for trajectory data anonymization. In *IEEE 13th Internation Conference on Data Mining Workshops*, 2013.

[95] E. Quinn. Privacy and the new energy infrastructure. (09), 2008.

[96] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor. Smart meter privacy: A utility-privacy framework. *IEEE International Conference on Smart Grid Communications*, pages 190–195, 2011.

[97] S. Ramchurn and P. Vytelingum. Putting the'smarts' into the smart grid: a grand challenge for artificial intelligence. *Communications of the ACM*, 55(4):86–97, 2012.

[98] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. Jennings. Agent-Based Control for Decentralised Demand Side Management in the Smart Grid. pages 2–6, 2011.

[99] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the International Conference on Management of Data*, page 735, 2010.

[100] I. Richardson, M. Thomson, D. Infield, and C. Clifford. Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10):1878–1887, 2010.

[101] G. Sáenz de Miera, P. del Río González, and I. Vizcaíno. Analysing the impact of renewable electricity support schemes on power prices: The case of wind electricity in Spain. *Energy Policy*, 36(9):3345–3359, 2008.

[102] F. Schweppe, M. Caramanis, and R. Tabors. Evaluation of Spot Price Based Electricity Rates. *IEEE Transactions on Power Apparatus and Systems*, PAS-104(7):1644–1655, 1985.

[103] P. Scott, S. Thiébaux, M. van den Briel, and P. Van Hentenryck. *Principles and Practice of Constraint Programming*, volume 8124. 2013.

[104] S. Shafiee and E. Topal. When will fossil fuel reserves be diminished? *Energy Policy*, 37(1): 181–189, 2009.

[105] L. Shou, X. Shang, K. Chen, G. Chen, and C. Zhang. Supporting Pattern-Preserving Anonymization For Time-Series Data. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–16, 2011.

[106] Statistisches Bundesamt. *Statistisches Jahrbuch.* 2012. URL `http://www.destatis.de`.

[107] L. Sweeney. Guaranteeing anonymity when sharing medical data, the Datafly System. *AMIA Annual Fall Symposium*, pages 51–5, 1997.

[108] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[109] Transnet BW. Statistical PV and CHP data, 2012. URL `http://transnet-bw.de/eeg-and-kwk-g/eeg-anlagendaten/`.

[110] US Legislation. Annotation 19 - First Amendment.

[111] R. van Gerwen, S. Jaarsma, and R. Wilhite. Smart Metering. *Distributed Generation*, (July): 1–9, 2006. URL `www.leonardo-energy.org`.

[112] D. Varodayan and A. Khisti. Smart meter privacy using a rechargeable battery: Minimizing the rate of information leakage. *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1932–1935, 2011.

[113] A. Wagner, S. Speiser, O. Raabe, and A. Harth. Linked Data for a Privacy-Aware Smart Grid. *Lecture Notes in Informatics*, 1, 2010.

[114] Q. Wang, V. Megalooikonomou, and C. Faloutsos. Time series analysis with multiple resolutions. *Information Systems*, 35(1):56–74, 2010.

[115] T. Warren Liao. Clustering of time series data — a survey. *Pattern Recognition*, 38(11): 1857–1874, 2005.

[116] C. Weinhardt, C. Holtmann, and D. Neumann. Market-Engineering. *Wirtschaftsinformatik*, 45(6):635–640, 2003.

[117] M. Wong. *Discrete Fourier Analysis*. 2011.

[118] X. Xiao, G. Wang, and J. Gehrke. Differential Privacy via Wavelet Transforms. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1200–1214, 2011.

[119] E.-h. Yang and J. C. Kieffer. Efficient Universal Lossless Data Compression Algorithms Based on a Greedy Sequential Grammar Transform. 46(3):755–777, 2000.

[120] R. Yarovoy and F. Bonchi. Anonymizing moving objects: how to hide a MOB in a crowd? *Proceedings of the 12th International Conference on Extending Database Technology*, pages 72–83, 2009.

[121] M. Zeifman and K. Roth. Nonintrusive Appliance Load Monitoring: Review and outlook. *IEEE Transactions Consumer Electronics*, 57(1):76–84, 2011.

[122] J. Zoellner, P. Schweizer-Ries, and C. Wemheuer. Public acceptance of renewable energies: Results from case studies in Germany. *Energy Policy*, 36(11):4136–4141, 2008.

Index