Pers Ubiquit Comput (2013) 17:653–662
DOI 10.1007/s00779-012-0513-6

ORIGINAL ARTICLE

Re-identification of Smart Meter data

Erik Buchmann · Klemens Böhm · Thorben Burghardt · Stephan Kessler

Received: 29 September 2011 / Accepted: 1 February 2012 / Published online: 30 March 2012

Abstract The Smart Grid approach enhances the power grid with information technology. Smart Meters are an important part of the Smart Grid. They record the energy consumption of households with a high-resolution and transfer consumption records to the energy provider in real time. Since they allow to infer personal information like the daily routine of the household members, Smart Meters are also a promising source for lifelogging. However, in liberalized energy markets, many different parties have access to these data. This puts the privacy of consumers at risk. In this paper, we analyze to which degree Smart Meter data, as collected by our industry partner, can be linked to its producer, using simple statistical measures. We devise features of the energy consumption, for example, the first peak of demand in the morning, and we describe an analytical framework that quantifies how well these features can identify households. Finally, we conduct a study with 60,480 energy-consumption records from 180 households. Our study shows that 68 % of the records can be re-identified with simple means already. This insight is important for Smart Grids, as it emphasizes the need for research and use of anonymization techniques for the Smart Grid.

E. Buchmann (✉) · K. Böhm · S. Kessler
Karlsruhe Institute of Technology, Am Fasanengarten 5,
76131 Karlsruhe, Germany
e-mail: erik.buchmann@kit.edu

K. Böhm
e-mail: klemens.boehm@kit.edu

S. Kessler
e-mail: stephan.kessler@kit.edu

T. Burghardt
Lucebit GmbH, Konrad-Zuse-Ring 6, 68163 Mannheim,
Germany
e-mail: thorben.burghardt@lucebit.com

Keywords Smart grid · Re-identification · Privacy threats

1 Introduction

The Smart Grid is an initiative to save energy, to increase the utilization of electricity sources, to foster renewable energy and to integrate services like e-mobility or demand-side management into the electricity-distribution infrastructure. The Smart Grid builds upon information technology to manage a huge number of energy sources and consumers. Smart Meters are an important component of the Smart Grid. They record the energy consumption of households with high precision and high frequency, and they allow for automated and remote meter readings over data networks. Smart Meters also increase the transparency of the energy consumption. This gives way to a flexible pricing that encourages consumers to reduce peak demand and to shift usage to off-peak hours. Smart Meters also support precise load forecasts, smart appliances and services like energy consultations or demand-side management.

On the other hand, the time series produced by a Smart Meter allow to infer personal details [16], for example, the daily routine or the presence of specific electrical devices in a household (cf. Fig. 1). Recent studies have shown that it is even possible to use data from off-the-shelf Smart Meters to identify the TV movies viewed [10]. This is because unique fluctuations in the brightness of the movies have a measurable effect on the energy consumption of the TV set. Thus, Smart Meters are not only important to save energy. They also produce and distribute streams of energy-consumption data that might allow lifelogging at an amazing level of detail. Lifelogging [8, 15] refers to various devices and sensors used by individuals that record,

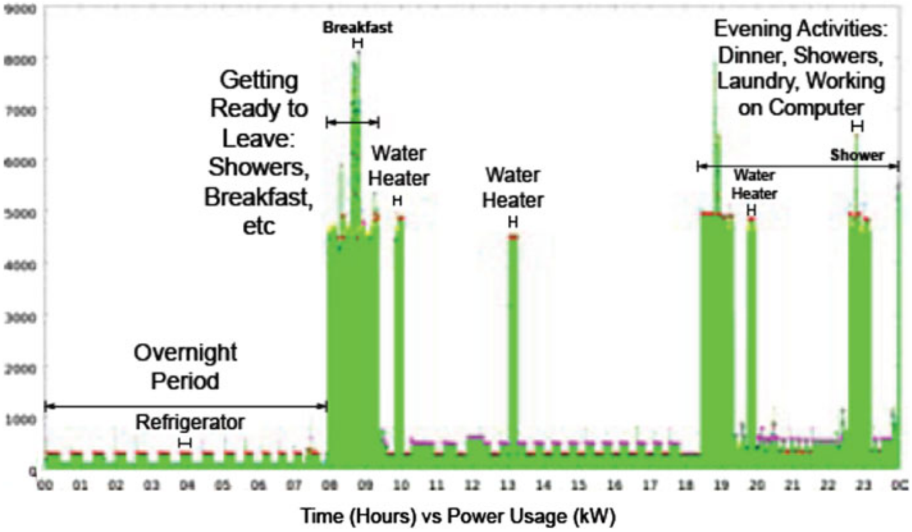

Fig. 1 Example of smart Meter data (reprinted from [16] with permission of the authors)

store and publish data like first-person video streams, physiological data or GPS coordinates. But while classical lifelogging is a deliberate act of an individual who uses wearable computers and other information technology to record her daily life [8, 13, 15], Smart Meters generate and distribute personal data without such a deliberate decision, with unpredictable consequences for the privacy of the individual.

Example 1 A network operator works with energy-consumption data from a certain area. Since the operator is not involved in billing or cashing, it does not know the identity of the households the data comes from. However, an employee of the operator knows that his neighbor typically uses the coffee machine at 7.15 am and the microwave at 1.30 pm. Suppose that only one time series from the consumption data has these characteristics. In this case, the employee could find out which consumption data belong to his neighbor and explore his entire consumption history.

The example illustrates *re-identification*, which is the process of assigning personal data without any personal identifier with the identity of its owner. Re-identification and inferring personal information are orthogonal to each other. Even if it is impossible to construct a quasi-identifier (i.e., a pseudonym) from the Smart Meter data, it might still be possible to extract sensitive information, and vice versa. For example, consider a set of identical time series of Smart Meter data from different persons. It is not possible to compute pseudonyms that distinguish the time series, but information such as the daily routine can be extracted.

Re-identification of Smart Meter data is a severe privacy threat: First, Smart Meters pervade the everyday life of many individuals. In Germany and parts of Australia, the installation of Smart Meters is required by law. In other countries, for example, Sweden, Italy or the United States, a lot of Smart Meters have been already installed without obligation. Second, the liberalization of the electricity markets implies that many different parties will have access to consumption data, for example, network operators, network carriers and electric utility providers. Thus, it will be necessary to publish Smart Meter data. Note that the privacy threats arising from Smart Meter data are much more severe if re-identification and extraction of sensitive information are possible at the same time. In particular, the impact of personal data would be much lower, if it is impossible to infer highly sensitive information or to assign such data with an identity.

In this paper, we analyze to which extent anonymous energy-consumption records, as collected by our industry partner, are prone to re-identification. 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 are important for business and research, as they point out in which way privacy obligations apply to Smart Meter data stripped from personal identifiers.

We have conducted a case study with 60,480 Smart Meter readings, which are similar to Fig. 1. Our data have been collected from 180 households with a metering frequency of one hour over a period of 14 days. 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. Since the daily routine is influenced by many aspects of the household, for example, employment status, hobbies or the number of persons, features of the energy-consumption data should be inherently identifying for many households. We consider features like the aggregated consumption per day or the time of the first peak demand in the morning, and we analyze to which extent we can use these features for re-identification. Instead of striving for very large sets of features or sophisticated algorithms for re-identification, we are interested in finding out if rather elementary features and relatively simple statistical measures are sufficient for re-identification of energy-consumption data. If so, the potential for misuse of Smart Meter data is high. Non-experts would be able to perform re-identification with simple measures. Straightforward feature values that could be guessed or observed by an adversary would increase the privacy threat even more. Finally, from a legal perspective, Smart Meter data must be treated with special care in many countries if the effort of re-identification is low.

In this paper, we make the following contributions:

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

Our study shows that 68 % of all consumption data can be re-identified, that is, we have found unique combinations of feature values in the energy-consumption data of 122 households. This shows that the privacy of a large share of households is at risk, because Smart Meter data might be personally identifiable lifelogging data. The most identifying features are the 0.9-quantile, the consumption on weekdays from 4:00 am to 8:00 am and the frequency of the most frequent consumption value. Our findings stress the need for future research on anonymization, perturbation or information-hiding techniques in the context of the Smart Grid. Our results are also very important for our industry partner and other companies in the field: Smart Meter data must be processed and handled according to privacy standards of society and data privacy laws, even if it does not contain personal identifiers.

Paper structure: Sect. 2 presents related work. Section 3 describes our study, followed by a description of the results in Sect. 4. We discuss our results in Sect. 5 and conclude in Sect. 6.

2 Background

In this section, we explain the connection between Smart Meter technology and lifelogging. Furthermore, we outline how private information can be obtained from Smart Meter data, and we describe re-identification approaches from research.

2.1 Lifelogging and Smart Meters

In general, lifelogging [8, 15] means using various devices and sensors to record, store and publish data like first-person video streams, physiological data or GPS coordinates. Such data can be used to support people with dementia, as a personal diary, for medical diagnoses and many other use cases. Recently, the use cases of lifelogging have been extended toward social experience [2, 20] by using Blogspot[1], Twitter[2], Facebook[3] and similar social media.

Some current lifelogging approaches [2, 13] make secondary use of existing appliances, for example, smartphones, PDAs, webcams or RFID tags ("passive lifelogging" [19]). In particular, these lifelogging approaches try to extract the user context, events of the daily routine and similar information from such data sources. Thus, Smart Meters that record the energy consumption of a household with a fine-grained level of detail can be used for passive lifelogging.

Data mining on Smart Meter data has the potential to reveal lifelogging information [11, 16]. This is because many daily activities involve the usage of electrical devices. The metering capabilities of Smart Meters vary widely. Basic models measure the energy consumption of a household with a temporal resolution of one hour. Sophisticated meters in turn have a sampling frequency in the range of milliseconds and also measure values like phase shift or reactive power consumption. High-resolution data allow to recognize individual electrical devices [11]. Machine learning approaches, for example, clustering, Bayesian networks or classification, give way to the extraction of personal information, even with low-resolution data [16]. Examples of such personal information are the number of individuals being at home, the daily routine or the employment status. Early approaches such as [11] have included a training phase to learn the consumption signature of electrical applications. Recent approaches extract usage patterns of electrical devices without training [12, 16]. Furthermore, studies have shown that data from off-the-shelf Smart Meters are sufficient to identify the TV movies viewed [10]. This is because unique fluctuations in the brightness of the movies influence the energy consumption of the TV set.

Privacy approaches for lifelogging, for example, for video streams [4], audio streams [22], recorded

[1] http://blogspot.com.

[2] http://twitter.com.

[3] http://www.facebook.com.

conversations [5] or timestamped data objects [20], usually assume that the individual concerned can decide which data are published. Others argue that various privacy issues of lifelogging are unreasonably overstated [19], since with the advent of social media like Facebook or Twitter, people disclose personal details freely. However, Smart Meter data are transferred automatically to others, that is, without the individual concerned observing which personal information is disclosed. In consequence, privacy approaches for Smart Meter data must be different from ones tailored for lifelogging.

2.2 Privacy threats of Smart Meter data

One might assume that Smart Meter data are anonymous data, provided that identifiers like address of the household or serial number of the Smart Meter are removed. However, there is evidence that personal data records without identifiers can be linked to individuals, based on external knowledge. For example, 63 % of the US citizens can be identified by the combination of the date of birth, gender and ZIP code [9]. Such a combination of attributes is called a quasi-identifier. In 2002, it has been demonstrated [21] that these quasi-identifiers allow to re-identify medical records by using the voter list as external knowledge. At this time, both data sets were publicly available in the United States, and medical records were assumed to be anonymous. Other examples include the re-identification of US census records [6] or AOL search engine records [1]. In consequence, re-identification might be also possible for Smart Meter data, provided that we can identify features of the energy consumption which can serve as quasi-identifiers. This is a severe privacy threat: The liberalization of the energy market requires to publish Smart Meter data to many different parties with different external knowledge. Even if such data are published without personal identifiers, re-identification might allow to compute quasi-identifiers for the individuals concerned. The quasi-identifier might serve as a unique fingerprint of an individual. This allows to interlink the databases of many stakeholders of the Smart Grid, even if some of them do not know names or addresses of the individuals. If the quasi-identifiers are known, it is possible to anonymize data sets to prevent re-identification. Related anonymization approaches have already been investigated, for example, in the fields of relational databases [14, 18, 21] or GPS trajectories [17].

The privacy problems described so far might affect society, for three reasons. Firstly, Smart Meters already have arrived at the mass market. In Italy, the energy provider Enel[4] has equipped 32 million of its customers with Smart Meters. In Germany, Smart Meters are required by law for any new or reconstructed building. Initiatives to install Smart Meters also exist in Sweden, Canada, the Netherlands and other countries. Secondly, the Smart Grid as such is still under development, and it has been designed to foster innovations toward energy efficiency with as few regulatory limitations as possible. Thus, at this moment, it is unclear which institutions will come into existence in the future, and which kinds of external knowledge they will be able to link to Smart Meter data. In the third place, since the Smart Grid is designed to give way to liberalized energy markets, the number of stakeholders that require access to Smart Meter data will be large. With many individuals having access to Smart Meter data, it is very unlikely that security measures like access control are suitable to protect the privacy of the individuals concerned. Our study is intended to acknowledge that Smart Meter data without personal identifiers can be identifying. Since such data might carry lifelogging information, this must be considered when developing new business cases, energy markets or energy services. Furthermore, since we analyze identifying features of the energy consumption, this paper is an important step toward anonymization of Smart Meter data.

3 Study methodology

In this section, we describe our study setup, the energy-consumption features considered, and an algebraic framework that measures to which extent the features allow for re-identification.

3.1 Study overview

In order to find out to which degree energy-consumption data are identifying, we analyze 60,480 Smart Meter readings from 180 households, measured with a sampling rate of one hour over a period of two weeks. It is difficult to obtain large samples of Smart Meter data, because Smart Meters are just about to enter the mass market, and its data are subject to various regulations, for example, data privacy and consumer-protection laws. However, if we can show that many households can be re-identified from this short time period, we have demonstrated that Smart Meter data are a serious privacy threat. The study data do not contain personal identifiers, and the profiles of our study households are similar to each other. Thus, re-identification is not as simple as distinguishing the consumption patterns of, say, unemployed individuals, shift workers or daytime employees. Instead, we have a challenging setup where re-identification requires an analytical framework that

[4] http://www.enel.com.

considers combinations of consumption features. Our framework performs re-identification in three steps:

Step 1: Feature and distance computation First we compute feature values of the energy consumption. For example, the value corresponding to the feature “Average Wakeup Hour” is the average time of the first increase of energy consumption in the morning. Furthermore, we calculate distances for the features of different data sets to identify similar set of feature values.

Step 2: Weights computation In this step, we assign weights to the features. Features with a high spread of values are likely to facilitate a differentiation of the households and should have higher weights. We use three different approaches (Static, Linear Optimization and Integer Linear Optimization) to determine weights. We are particularly interested to find out if computationally expensive optimizations increase re-identification performance.

Step 3: Re-identification We perform re-identification by calculating the weighted distance between the features of an anonymous consumption record and the ones of a household. A record is re-identified if its distance to the correct household is smaller than the distance to any other household.

3.2 Consumption features

We assume that the daily routine influences the energy-consumption data. We will use features to match the consumption data with the households. Features are suitable for an adversary to re-identify households if they have the following properties:

- It is difficult to change features of the energy consumption without changing the daily routine. For example, “Average Wakeup Hour” is a promising feature in this sense.
- It is possible for an adversary to guess feature values just by observing the way of life of an individual.
- Consumption-feature values of a household at different times should be similar to each other, but different from feature values of other households at any time.

Table 1 shows the features which we have tested for our study. Our features are based on full weeks of metering data. This is because the life of most individuals follows a weekly pattern. We distinguish three classes of features. The first class (first row of Table 1) considers features that are solely based on the energy consumption. “Overall Consumption” is the total energy consumption over one week. “Maximum Consumption” and “Minimum Consumption” are the respective maximum or minimum of the consumption during one week. “0.9-Quantile” is the value that is larger than or equal to 90 % and smaller than or equal to 10 % of all consumption values. “Frequency of mode” is the number of the most frequent consumption values. “Standard deviation” is the standard deviation of all consumption values over a week of our test data. The second class (second row of Table 1) includes features that consider both time and consumption. “Consumption Mo-Fr 4 am–8 am”, “Consumption Mo-Fr 10 am–4 pm” and “Consumption Mo-Fr 9 pm–2 am” are the aggregated consumption values over one week during breakfast, lunch and night time. “Weekend Consumption” is the sum of the consumption over the weekend. The third class (third row of Table 1) is based on time. “Average Wakeup Hour” is the average time of the first increase of energy usage in the morning, usually caused by making breakfast or other typical morning activities. Accordingly, “Average Bedtime Hour” is the point in time where the energy consumption decreases at night. Note that the list of features is necessarily incomplete. But this is not a problem. Instead of striving for completeness, our goal is to analyze to which degree energy-consumption data are identifying when using simple means.

Table 1 Energy-consumption features

Absolute difference	Relative difference
Consumption	
Overall consumption	Maximum consumption
Minimum consumption	Standard deviation
0.9-Quantile	Frequency of mode
Consumption during time interval	
Consumption Mo-Fr 4 am–8 am	Consumption Mo-Fr 10 am–4 pm
Weekend consumption	Consumption Mo-Fr 9 pm–2 am
Time	
Average wakeup hour	
Average bedtime hour	

To re-identify a consumption record, we measure the distances between feature values of known households and feature values calculated from the record in question. We consider both absolute and relative distances (first and second column of Fig. 1). Our motivation is that in some cases, relative distances might be more informative than absolute ones. For example, switching a light bulb on or off may be as important for re-identification as turning a coffee machine on or off, even if the coffee machine has a much higher energy consumption.

3.3 Algebraic framework

We now describe the algebraic framework we have used for our study. The framework computes a value for each feature. To decide whether two sets of feature values are from the same household, we compute the weighted distance between them. We make use of three different alternative methods to determine the weight of a feature, in order to compare these methods.

Let H be a set of households $h \in H$, and let m_h^z denote a time series of energy-consumption records of h in time interval z. The time series consist of records containing a time stamp and the values measured. For our study, we distinguish two distinct time intervals: training period θ and re-identification period ρ. A function $v_f(m_h^z)$ computes the value of the feature f with $f \in$ {*overall consumption*, ..., *average bedtime hour*}. The goal of our framework is to decide whether a data set from the training period $m_{h'}^{\theta}$ and a data set from the re-identification period $m_{h''}^{\rho}$ are from the same household, according to the similarity of their feature values.

Step 1: Feature and distance computation We start by calculating values of the features given in Table 1, i.e., we calculate $v_f(m_h^z)$. Furthermore, we calculate the absolute and the relative distance $d_f^{\mathrm{abs}}(m_i^{\theta}, m_j^{\rho})$, $d_f^{\mathrm{rel}}(m_i^{\theta}, m_j^{\rho})$ of two consumption records m_i^{θ}, m_j^{ρ} according to feature f as follows:

$$d_f^{\mathrm{abs}}(m_i^{\theta}, m_j^{\rho}) = \left| v_f(m_i^{\theta}) - v_f(m_j^{\rho}) \right|$$

$$d_f^{\mathrm{rel}}(m_i^{\theta}, m_j^{\rho}) = \left| \frac{2(v_f(m_i^{\theta}) - v_f(m_j^{\rho}))}{v_f(m_i^{\theta}) + v_f(m_j^{\rho})} \right|$$

To re-identify a consumption record, we compare feature values of known households with values calculated from the record in question. We assume that two features are similar if the difference between them is under a threshold. In order to compute this threshold, we first define a set D_f of differences between the feature values from the training period and the re-identification period of each household $h \in H$ with $n = |H|$: $D_f = \{d_f(m_{h1}^{\theta}, m_{h1}^{\rho}), ..., d_f(m_{hn}^{\theta}, m_{hn}^{\rho})\}$. This set might contain outliers. For example, if a person wakes up hours earlier than usual in the training period, the feature "Average Wakeup Hour" for this household contains an outlier. To diminish the influence of outliers on the threshold, we define a set $D_f^{'}$ containing the smallest 90 % of the distances between feature values in D_f.

Thus, $D_f^{'} \subset D_f$, $|D_f^{'}| = 0.9 \cdot |D_f|$, and $\forall\ e \in (D_f - D_f^{'})$: $e \geq \max\ (D_f^{'})$. Based on $D_f^{'}$, we now define the threshold T_f for each feature: $T_f = \mathrm{avg}\ (D_f^{'}) + \mathrm{SD}\ (D_f^{'})$.

Finally, we compute a normalized feature score for each feature, based on the threshold T_f and the standard deviation of the differences in $D_f^{'}$. The more similar two features are, the smaller is this score.

$$\mathrm{Score}_f(m_i^{\theta}, m_j^{\rho}) = \begin{cases} 0 \text{ if} \quad d_f(m_i^{\theta}, m_j^{\rho}) < T_f \\ \frac{d_f(m_i^{\theta}, m_j^{\rho}) - T_f}{\mathrm{SD}(\mathrm{D}_\mathrm{f}')} \quad \text{otherwise} \end{cases}$$

Step 2: Weights computation The score of two consumption records is the sum of all normalized feature scores over all features. We calculate the score as follows:

$$\mathrm{Score}(m_i^{\theta}, m_j^{\rho}) = \sum_{f \in F} w_f \cdot \mathrm{score}_f(m_i^{\theta}, m_j^{\rho})$$

Our goal is to re-identify a consumption record m_h^{ρ} from the re-identification period ρ by using the feature values from the training period θ as external knowledge. Therefore, we calculate the score of this record m_h^{ρ} and the training records m_i^{θ} of all households $i \in H$. Our Score function makes use of weights, which should be high for distinct features and low for less-distinctive ones. We now define three different approaches to determine these weights. Our baseline is the **Static Approach**, where all weights are set to 1:

$$\forall f \in F : w_f = 1$$

The **LP Approach** uses Linear Optimization to determine a set of weights that maximize the difference between incorrectly and correctly re-identified consumption data. We use Linear Optimization to maximize a term that iterates over each household $i \in H$ and sums the differences between the score of a correctly re-identified i and the next closest household j:

$$\sum_{i \in H} \left| \min_{\substack{j \in H \\ i \neq j}} (\mathrm{Score}(m_i^{\theta}, m_j^{\rho})) - \mathrm{Score}(m_i^{\theta}, m_i^{\rho}) \right|$$

The **ILP Approach** is based on Integer Linear Optimization with a binary variable $x_i \in \{0, 1\}$ and $i, j \in H, i \neq j$:

$$x_i = \begin{cases} 1 & \text{if } \exists j : \mathrm{Score}(m_i^{\theta}, m_i^{\rho}) \geq \mathrm{Score}(m_i^{\theta}, m_j^{\rho}) \\ 0 & \text{otherwise} \end{cases}$$

The variable x_i is 0 if the score for a correct re-identification is smaller than for an incorrect one, and 1 otherwise. We use Integer Linear Optimization to determine weights that minimize the following term:

$$\sum_{i \in H} x_i$$

Step 3: Re-identification Finally, we use our framework for re-identification. A consumption record has been re-identified if the score from training period and re-identification period of the same household is smaller than the score of different households, that is, if $\forall i \neq h, i \in H : \text{Score}(m_h^\theta, m_h^\rho) < \text{Score}(m_i^\theta, m_h^\rho)$.

To re-identify each time series, we conduct a fivefold cross validation: We partition our data into five partitions, each partition containing 20 % of the households. We use four partitions as training data to compute the weights and the thresholds, and we re-identify the remaining partition. We repeat this process five times, so that each partition is used for training as well as for re-identification. At the end, we compute the average of the results.

4 Study results

In this section, we show the results of our case study with 60,480 Smart Meter readings. We first describe some details of our features, followed by an analysis of the weights we have obtained. Finally, we determine to which degree it is possible to re-identify Smart Meter data.

4.1 Features

A feature (cf. Table 1) suited for re-identification has (1) a large spread in its absolute values and (2) small differences between the feature values of the same household at different points in time. For illustrative purposes, we analyze the feature “Standard deviation”. This feature computes the standard deviation over all values measured during a time interval.

Figure 2 shows the histogram of the standard deviation of our study data. The standard deviation varies between 0 and 0.6. Most of the households have a standard deviation smaller than 0.4. Recall that our framework normalizes the feature values. At most 12 of our 180 households have the same standard deviation. Thus, Fig. 2 confirms that the feature “Standard deviation” has a high spread of values.

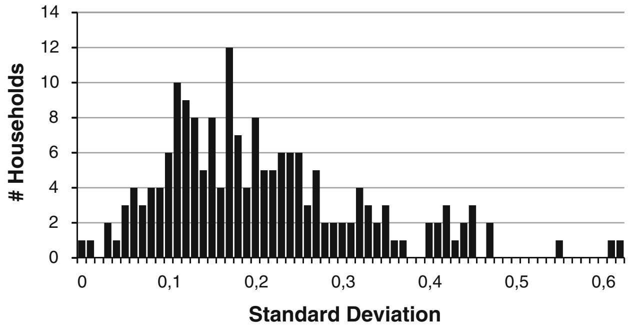

Fig. 2 Feature “standard deviation”

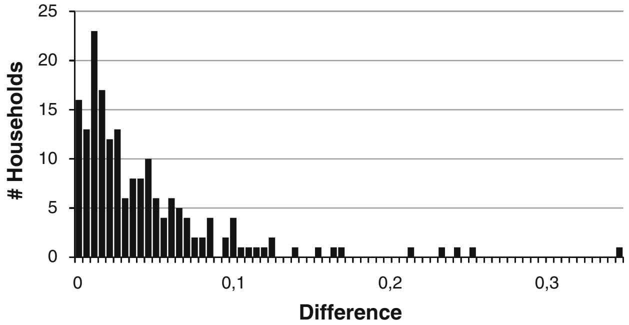

Fig. 3 Differences between values of feature “Standard deviation”

Figure 3 shows a histogram of the differences between the first and the second week of the feature “Standard deviation” for the same household. Except for some outliers, most households have a difference smaller than 0.1. Thus, this feature characterizes a household rather well. Figure 4 shows a histogram of the relative differences for the same feature. The figure indicates that the feature values of the same household do not differ much with the relative distance measure. Since the spread of values (cf. Fig. 2) is more distinctive for this feature than absolute values, we use the relative distance measure for feature “Standard deviation”.

The first column of Table 2 lists the standard deviation of all features we consider. The first column indicates that some features have a very large spread of values, for example, “Overall Consumption” or “Frequency of Mode”. The second column of Table 2 shows the standard deviation of the differences between feature values from the first and the second week of the same household. Features with a high standard deviation and a low difference promise to re-identify households very well.

4.2 Weights

Our framework uses the weights w_f to reflect the utility of a feature f for re*identification. Features that better distinguish one household from another one have a higher

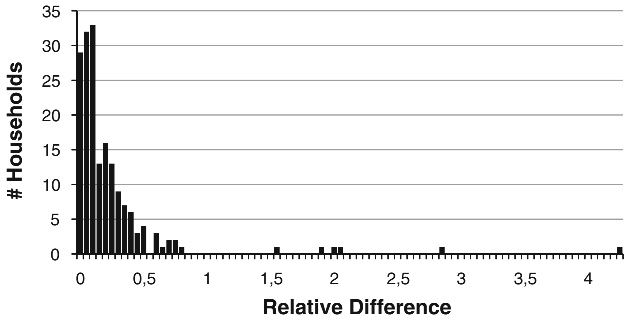

Fig. 4 Relative differences between values, Feature “Standard deviation”

Table 2 Standard deviation of feature values and differences

Standard deviation of	Feature values	Differences
Overall consumption	23.04	6.55
Minimum consumption	0.045	0.020
Maximum consumption	0.76	0.40
Standard deviation	0.12	0.051
Mo-Fr 4-8	2.32	1.01
Mo-Fr 10-16	5.73	2.16
Mo-Fr 21-2	4.25	1.78
Weekend consumption	0.068	0.056
Wakeup hour	3.55	1.69
Bedtime	2.48	1.54
0.9-quantile	0.27	0.11
Frequency of mode	18.85	9.82

Table 3 Feature weights

Weights	ILP	LP	Static
Overall consumption	0.48	0.01	1.00
Minimum consumption	1.20	0.96	1.00
Maximum consumption	0.12	0.72	1.00
Standard deviation	1.08	0.96	1.00
Mo-Fr 4-8	1.80	0.96	1.00
Mo-Fr 10-16	1.20	1.44	1.00
Mo-Fr 21-2	0.96	1.44	1.00
Weekend consumption	0.12	0.36	1.00
Wakeup hour	0.84	0.48	1.00
Bedtime	0.72	0.48	1.00
0.9-quantile	1.92	0.96	1.00
Frequency of mode	1.56	2.16	1.00
Sum	12	12	12

weight than others. Table 3 lists the weights of all features. The features “0.9-Quantile”, “Mo-Fr 4–8”, “Mo-Fr 10–16”, “Minimum Consumption” and “Frequency of Mode” have the highest weights. “0.9-Quantile” reflects the energy consumption when many electrical devices are used at the same time. “Mo-Fr 4–8” and “Mo-Fr 10–16” refer to the living habits of the household in the morning and at noon. “Minimum Consumption” reflects the energy consumption of devices that are always on, even when all members of the household are out or asleep. Examples are refrigerators or devices on standby. Thus, this feature characterizes a household well and does not change much over the time period of our study. “Frequency of Mode” is the number of times the most frequent consumption is measured. The most frequent consumption value usually is near “Minimum Consumption”. Features like “Weekend Consumption”, “Bedtime” or “Overall Consumption” fluctuate much, that is, they do not show strong correlations to regular habits or to the devices used. Thus, such features are assigned to a smaller weight. ILP and LP produce different weights, but show similar tendencies. For example, Frequency of Mode has the highest weight with both approaches, while features like Weekend Consumption are weighted down.

4.3 Re-identification

We now analyze the re-identification performance of our framework. The feature weights are computed over a set of households disjoint from the households to be re-identified. Thus, re-identification only requires the feature values of a household.

Table 4 lists the number of correct re-identifications. Each column shows the absolute number of re-identified records per approach, as well as the percentage. Each of the five rows represents a different data partition from the fivefold cross validation. The last row contains the average of all partitions. Table 4 shows that ILP produces the highest number of correct re-identifications (68.3 %). LP (64.4 %) does not perform as good as ILP, but is still better than using static weights (63.3 %).

In order to find out how many re-identifications were close-by, we have counted the data sets where the originator is within the top-3 households with the smallest score. In this case, we have not performed re-identification, but we have reduced the uncertainty to one out of three households. Table 5 displays the results of this test. The table shows that ILP is still in front with 82.8 %. The difference between LP (81.1 %) and static weights (79.4 %) has increased slightly.

In general, the re-identification rate of the static approach is much higher than we had expected (63.3 %). This indicates that we have intuitively selected our features very well, and re-identification is possible just by using simple statistics and intuitively selecting adequate consumption features. The difference between static weights and weights determined by LP or ILP is rather small, and both approaches tend to assign similar weights. We

Table 4 Re-identification performance

Fivefold cross val.	ILP		LP		Static	
	abs	%	abs	%	abs	%
1	24	66.7	23	63.9	22	61.1
2	26	72.2	25	69.4	23	63.9
3	25	69.4	23	63.9	24	66.7
4	24	66.7	21	58.3	23	63.9
5	24	66.7	24	66.7	22	61.1
avg	24.6	68.3	23.2	64.4	22.8	63.3

Table 5 Close-by re-identifications (Top-3)

Fivefold cross val.	ILP		LP		Static	
	abs	%	abs	%	abs	%
1	31	86.1	30	83.3	30	83.3
2	30	83.3	27	75.0	27	75.0
3	29	80.6	29	80.6	29	80.6
4	29	80.6	28	77.8	27	75.0
5	30	83.3	32	88.9	30	83.3
avg	29.8	82.8	29.2	81.1	28.6	79.4

conclude that selecting appropriate features is more important than spending much computational effort for computing optimal weights. However, we expect that weights would become more important in cases where a larger number of features is used for re-identification.

5 Discussion

The goal of our study has been to find out to which degree Smart Meter data can be re-identified by simple means, that is, based on educated guesses regarding certain features of the energy consumption and simple statistical measures. Educated guesses can be obtained from various sources. For example, a person could observe that her neighbor never is at home on Saturday evenings and hear the use of appliances, for example, laundry machine or TV set. Thus, the person knows at which times nobody does switch devices on or off, and she knows when devices with a well-known energy-consumption profile (cf. Fig. 1) have been used. Many other sources of information might allow to guess feature values. For example, data about the working periods of shift workers correspond to features like Wakeup Hour or Weekend Consumption.

Our approach re-identifies 68.3 % of the data. Further, we have seen that it does not require much data to derive feature values for re-identification. For example, the share of correctly re-identified data does not decrease much if we skip the feature values with the lowest weights. This also implies that an attacker would be able to re-identify energy-consumption data just by guessing the values of a small number of features of a household. Thus, our study has demonstrated that Smart Meter data are inherently identifying. This is relevant when it comes to the processing and dissemination of such data. In order to protect the privacy of energy consumers, it is not sufficient to simply remove identifiers from the data streams generated by Smart Meters. We expect that even short time intervals of consumption data would be sufficient for re-identification. Thus, it appears promising to apply perturbation or information-hiding techniques before Smart Meter data are transferred to others.

We expect that (1) more sophisticated features and (2) improvements of the algebraic framework can increase the share of re-identified data. More specifically, one optimization could be to determine features according to the data available. Thus, if metering data from Friday were not available, the feature “Mo-Fr 4–8” could be changed to “Mo-Thu 4–8”. Another optimization could try to determine features automatically. Since a time series of Smart Meter data are similar to a feature vector in multimedia retrieval or high-dimensional data mining, feature-selection algorithms from these areas of research might be applicable to determine optimal features.

The algebraic framework could implement state-of-the-art algorithms from data mining for outlier-detection. Currently, our framework simply ignores the largest 10 % of the distances between feature values. Another optimization could extend the thresholding for the similarity of feature values. Our framework uses the sum of the average and the standard deviation of the distances between the feature values. A more elaborate approach could use genetic algorithms or other approaches to solve optimization problems to obtain a threshold that increases the number of correctly re-identified data sets.

Note that optimizations showing that it is possible to re-identify an even larger share of Smart Meter data are orthogonal to the purpose of this study. Our objective was to show that the effort to re-identify Smart Meter data can be very low. This has various implications on the handling and storage of such information, be it for lifelogging or Smart Grid services. For example, German data privacy laws follow the principle of proportionality. That is, an information is assumed to be anonymous as long as the effort to link this information to the person concerned is disproportionally large in comparison with the sensitivity of the data [3]. Due to Directive 95/46/EC [7], similar regulations exist in all other European countries. Thus, it is sufficient to show that simple means of re-identification exist so that Smart Meter data are subject to EU data privacy regulations, with many obligations [7] regarding data transmissions to third parties, rights of the individuals to obtain information or deletion, etc.

6 Conclusions

The Smart Grid uses information technology to increase the energy efficiency of the electricity grid. An important component of the Smart Grid are Smart Meters, which measure the energy consumption of households with a high precision and frequency. Since Smart Meter data allow to infer details of the daily routine or the electrical devices used, they are a promising data source for lifelogging. However, at the same time, they put the privacy of consumers at risk.

Together with our industry partner, we have shown that identifying households based on their energy-consumption records is feasible with relatively simple means already. We have identified and analyzed 12 intuitive features of the energy consumption, for example, the consumption on weekdays or the first increase of energy-consumption in the morning. We have also devised an analytical framework that allows us to analyze to which extent such features can be used to re-identify consumption records. Finally, we have conducted a study with 60,480 energy-consumption records from 180 households, which have been metered with a frequency of one hour over a period of two weeks. Our study has shown that 68 % of the energy-consumption records can be re-identified. Thus, the study has provided evidence that in some cases even guessing feature values of the energy consumption can be sufficient for re-identification. This insight is important for industry, as it emphasizes the need for anonymization and perturbation techniques in the context of the Smart Grid.

References

1. Barbaro M, Zeller T (2006) A face is exposed for AOL searcher no. 4417749. New York Times, New York
2. Barbu C, Kröner A, Schneider M, Jacobs O (2009) Studying the functions of sharable digital memories. IADIS Int J WWW/Internet (IJWI) 7(1):44–62
3. Bundesrepublik Deutschland (2003) Bundesdatenschutzgesetz (BDSG). Bundesgesetzblatt I/2003 S.66
4. Chaudhari J, Cheung SS, Venkatesh MV (2007) Privacy protection for life-log video. In: Proceedings of the workshop on signal processing applications for public security and forensics
5. Cunningham S, Truta TM (2008) Protecting privacy in recorded conversations. In: Proceedings of the 1st international workshop on privacy and anonymity in the information society (PAIS'08)
6. Dalenius T (1986) Finding a needle in a haystack or identifying anonymous census record. J Off Stat 2(3):329–336
7. European parliament and the council of the European Union: (1995) Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Off J L 281, 11/23/1995, p. 31
8. Gemmell J, Bell G, Lueder R (2006) MyLifeBits: a personal database for everything. Commun ACM (CACM) 49
9. Golle P (2006) Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of the 5th workshop on privacy in the electronic society (WPES'06)
10. Greveler U, Justus B, Löhr D (2011) Hintergrund und experimentelle Ergebnisse zum Thema "Smart Meter und Datenschutz". Available at http://www.daprim.de
11. Hart GW (1992) Nonintrusive appliance load monitoring. Proc IEEE 80(12):1870–1891
12. Kim H, Marwah M, Arlitt M, Lyon G, Hand J (2011) Unsupervised disaggregation of low frequency power measurements. In: Proceedings of the 2011 SIAM international conference on data mining (SDM'11)
13. Kröner A, Schneider M, Mori J (2009) A framework for ubiquitous content sharing. IEEE Perv Comput 8:58–65
14. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M (2006) l-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd international conference on data engineering (ICDE'06)
15. Mann S (1997) Smart clothing: the wearable computer and wearcam. Pers Ubiquit Comput 1(1):21–27
16. Molina-Markham A, Shenoy P, Fu K, Cecchet E, Irwin D (2010) Private memoirs of a Smart Meter. In: Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building
17. Nergiz ME, Atzori M, Saygin Y (2008) Towards trajectory anonymization: a generalization-based approach. In: Proceedings of the SIGSPATIAL ACM GIS 2008 international workshop on security and privacy in GIS and LBS
18. Ninghui L, Tiancheng L, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd international conference on data engineering (ICDE'07)
19. O'Hara K, Tuffield M, Shadbolt N (2009) Lifelogging: privacy and empowerment with memories for life. IDIS 1(2):155–172
20. Rawassizadeh R, Tjoa A (2010) Securing shareable life-logs. In: Proceedings of the 2nd international conference on social computing (SocialCom'10)
21. Sweeney L (2002) k-Anonymity: a model for protecting privacy. Int J Uncertain Fuzz Knowl Based Syst 10(5):557–570
22. Yonezawa T, Okamoto N, Yamazoe H, Abe S, Hattori F, Hagita N (2011) Privacy protected life-context-aware alert by simplified sound spectrogram from microphone sensor. In: Proceedings of the 5th ACM workshop on context-awareness for self-managing systems