Information Sciences 370–371 (2016) 355–367

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A Technique to provide differential privacy for appliance usage in smart metering

Pedro Barbosa*, Andrey Brito, Hyggo Almeida

Computer and Systems Department, Federal University of Campina Grande, Brazil

ARTICLE INFO

Article history:
Received 12 September 2015
Revised 14 June 2016
Accepted 6 August 2016
Available online 8 August 2016

Keywords:
Differential privacy
Meter reading
Noise addition
Smart grids

ABSTRACT

Smart meters read electric usage and enable power providers to collect detailed consumption data from consumers. Based on this data, power providers can perform and improve many services such as differential tariffs and load monitoring. However, these readings also gather personal information that can be intrusive and threaten consumers' privacy. Consequently there is an urgent need to address how to protect consumers' privacy when using smart meter systems. We propose a lightweight approach for offering privacy using noise addition. Since the consumer behavior is very correlated with appliance usage, we measure the privacy level achieved by appliances through the state of the art in privacy (i.e., differential privacy model) and evaluate a filtering attack to eliminate the added noise. The utility is validated in a discussion regarding the smart meter benefits and an evaluation if they can still be provided when using our proposed approach.

1. Introduction

Smart Grids are systems that combine the traditional power grid with modern information technologies to enable a more efficient and robust grid. With these systems, power providers can monitor, analyze and control the network and communicate with the consumers to improve efficiency, reduce energy consumption and cost, and maximize the transparency and reliability of the energy supply chain.

Big data is one of the most discussed phrases in the Smart Grid industry today. In addition, big data analytics can help power providers to evaluate the areas within their networks that can be refined or improved and help to assess the business benefits to be achieved as result of Smart Grid investments. In this scenario, a key component on the consumers' side is the use of smart meters, which measure energy usage at a fine granularity and can monitor detailed information regarding the consumers' behavior. These devices may change the commercial relations between consumers and power providers. On the one hand, it creates opportunities for new services by power providers, on the other, it raises concerns about consumers' privacy [8]. Clearly, there is a trade-off between utility and privacy.

Fine-grained data of electricity usage naturally include personal and privacy-sensitive information regarding which appliances are active, if the house is empty at a certain time; when the inhabitants wake up, take a shower, or shut down the television; or even if certain appliances are no longer operating at desired levels of efficiency. With advanced power signature analysis tools such as the Non-Intrusive Appliance Load Monitoring (NIALM), the attacker can find out what types of

* Corresponding author.
E-mail addresses: pedroyossis@copin.ufcg.edu.br, pedroyossis@gmail.com (P. Barbosa), andrey@dsc.ufcg.edu.br (A. Brito), hyggo@dsc.ufcg.edu.br (H. Almeida).

http://dx.doi.org/10.1016/j.ins.2016.08.011
0020-0255/

Table 1
Comparison of homomorphic encryption (HE), rechargeable batteries (RB), and our noise addition (NA) approaches.

	HE [10,14,15]	RB [3,19,24,34]	NA
Low complexity		√	√
Scalability		√	√
Meters' independence		√	√
Low cost	√		√
Low environmental impact	√		√
Accuracy	√	√	

appliances are used at any time and learn more detailed information about customer's daily activities. Batra et al. [7] design some methodologies to identify the use of appliances through load profiles. If the NIALM algorithm is running remotely, the homeowners may not know that their behavior is being monitored and recorded.

In a traditional industrial setting, such behavior information is not in principle useful for a power provider. However, currently this type of information is of interest to many businesses that want to identify the profile of a potential consumer of their products and services. Therefore, disclosing such profiles and habits of consumers evokes issues about privacy. Clear rules are needed to protect consumers from misuse of their behavioral data and to avoid that Smart Grids become a new type of Big Brother [8]. Unfortunately, protection laws may take decades to be applied whereas smart meters are already operational.

Currently, only a handful of solutions exist to protect Smart Grid privacy; most of these solutions have been tied to specific technologies, affect the provision of some services by power providers, or are computationally complex. Moreover, existing solutions have not quantified the loss of benefit (utility) and the reached level of privacy that results from any such privacy approach. In this work we propose an approach which does not suffer from these issues.

The main contributions of this paper are: i) the proposal of a novel lightweight privacy approach by adding noise; ii) privacy validation for appliances according to the state of the art in privacy (*i.e.*, the differential privacy model) and to the evaluation of a filtering attack; and iii) utility validation with a discussion regarding the smart meter benefits and an evaluation if they can still be provided when using our proposed approach.

The rest of the paper is organized as follows. We discuss related work in Section 2 and summarize the problem statement in Section 3. We present our approach for privacy preserving in Section 4, its privacy validation in Section 5, and its utility validation in Section 6. We provide a filtering attack evaluation in Section 7 and a utility improvement in Section 8. We evaluate the performance of our approach and compare with other proposals in Section 9. Finally, we conclude the paper in Section 10.

2. Related work

Most of the privacy solutions (which also preserve utility) proposed in the literature are based on homomorphic encryption or rechargeable batteries. Homomorphic schemes [10,14,15] are promising mechanisms that may solve many privacy concerns. These solutions seek to support utility and privacy in different ways; however, they are computationally complex [21] (for example, $O(n^2)$ [15] and $O(n)$ [14], where n is the number of meters) and the meters are not independent, since a cooperation between them is necessary or there should be a third party for key or random secret sharing distribution. If one meter fails in any message exchange, the aggregation becomes impossible because it requires computations using all distributed keys or secret shares. Therefore, these solutions may raise fault and scalability issues when used in a wide area with a large number of meters.

Rechargeable batteries [3,19,24,34] between appliances and smart meters can help to reduce the privacy issues as the appliance signatures are no longer legible. Nevertheless, in addition to the signatures of the batteries still leaking information, it is hard to ignore the environmental effects and the costs of using batteries. McLaughlin et al. [24] stipulate that a lead-acid battery of 50 Ah which operates at 12 V may cost approximately $100 and to achieve a typical residential nominal voltage of 120 V it is required 10 of such batteries (aprox. $1,000). The lifetime of each battery is approximately two years. Therefore, these solutions are not low cost.

Table 1 presents a comparison of homomorphic encryption (HE), rechargeable batteries (RB) and our noise addition approach (NA). Besides the fact that our approach releases noisy counts, this does not invalidate the benefits of using metering data. In our models, the maximum allowed error is parameterized and this makes the obtained error in an aggregation operation to be controlled.

Other related research is the work of Wang et al. [32], which proposes an approach to mask the data using GMM (Gaussian Mixture Models), and the works of Bohli et al. [9] and He et al. [16], which propose approaches to mask the data using Gaussian noise. Noise addition is a promising and efficient technique, however, these approaches do not have formal models to calculate the amount of noise that should be added to guarantee desired privacy and utility levels. In fact, He et al. [16] argue that for real world system design, a proper trade-off between privacy protection and accuracy should be considered. Moreover, these proposed noise addition techniques do not provide differential privacy guarantees for appliances.

Sankar et al. [29] propose a framework for smart metering privacy that, similar to differential privacy, could be used to evaluate our approach. After studying this framework we concluded that mutual information (their privacy metric) does not always present expected results because timing is not critical. For example, when two appliance events are swapped, it changes (permutes) the signal in time but not its probability density function; hence, in this context, the mutual information yields no privacy improvement. A "distance function" such as mean-square error (their utility metric) does not always present expected results either because it measures utility for individual measurements, not for aggregated values such as billing (however, the power provider is mainly interested in aggregated values, as stated by the features presented in Section 6). Moreover, this framework does not have any guarantee that an appliance is indistinguishable in a profile. In differential privacy this is the main goal and is achieved by construction.

There are works [1,18] dealing with differential privacy for consumers in a group, making a consumer indistinguishable in, for example, a regional load monitoring. However, in our context, the privacy problem in smart metering is not anonymization of a consumer in a group because the consumer identity is crucial for many services such as billing. Liao et al. [23] also proposes a technique to achieve differential privacy, but it is necessary a concentrator (which is responsible for noise addition) between the power provider and smart meters, and if this concentrator is not trusted, the users' privacy may be violated. Our proposed approach follows a *Trust No One* philosophy and applies noise addition and differential privacy in the smart meters for making appliance usages and human behaviors indistinguishable in a consumption profile, which are the real sensitive information in our context.

3. Problem statement

Some motivations for usefulness to collect high-resolution data are the following: identification of non-technical losses (*e.g.*, power thefts), optimizations in load forecasting, billing and pricing services, and monitoring of energy quality scores. Further, such data may be used to facilitate and improve network management, reduce peak load, calibrate the load distribution, and many other uses [29]. Therefore, collection and dissemination of energy information are critical to the Smart Grid. However, regarding privacy, information of power consumption may be used for purposes that are not related to energy management, thereby making it potentially dangerous to individual privacy of consumers. In fact, this is one of the main reasons for the lack of mass roll-out of smart meters in many countries [20].

Differential privacy is a recent area of research that brings mathematical rigor to the problem of privacy-preserving analysis of data. Informally, in our context, the definition stipulates that any appliance should be indistinguishable in the disclosed masked consumption profile. Thus, an attacker cannot learn anything regarding appliances on the disclosed profile, even in the presence of any auxiliary information the attacker may have.

The problem statement of this work is the following: how to preserve the privacy of consumers but making the provision of certain services by the power provider still possible? Since the privacy is related with appliance usage, we aim to achieve differential privacy for appliances in metering data without affecting the aggregate data (*e.g.*, the total consumption in a region during an interval of time, or the total consumption of a specific consumer throughout a billing period).

4. Our masking approach

Regarding consumption measurements, our proposed masking approach focuses on:

- Enabling the calculation of the total consumption of a consumer over a period of time (*e.g.*, a month for billing);
- Enabling the calculation of the total consumption of all consumers in a region at a certain instant of time;
- Avoiding the measurement of the instantaneous consumption of an individual consumer at a certain instant of time.

Therefore, if each consumer sends a consumption measurement periodically to the power provider, it can organize these values as a matrix, where the sum of a row refers to the total consumption of a consumer throughout the time period, and the sum of a column refers to the total consumption of all consumers from this group at an instant of time. The rows of the matrix can be used for billing purposes and the columns can be used for load monitoring.

If a consumer wants privacy in a sensitive moment, we propose that, instead of sending real measurements, the smart meter sends masked measurements to the power provider in a way to do not affect the results of the aggregating operations. To exemplify the proposed approach in this paper, billing (sum of a row) as a base application is considered.

For billing purpose, if at each individual measurement the smart meter reads the consumption and adds a random number, at the end of a billing period the result will be:

$$\sum_{i=1}^{N} c_i \approx \sum_{i=1}^{N} (c_i + x_i)$$

where N is the total number of measurements, x_i is a random number generated from a probabilistic distribution and c_i is an individual consumption measurement.

Table 2
Analytical models obtained for different probability distributions.

Distribution	Model	Comments
Arcsin	$e_o \sim N\left(0,\ \frac{NX^2}{2}\right)$	X is the range of the original distribution
Laplace	$e_o \sim N(0, 2Nb^2)$	b is the scale parameter of the original distribution
Normal	$e_o \sim N\left(0,\ N\sigma_X^2\right)$	σ_X^2 is the variance of the original distribution
Uniform	$e_o \sim N\left(0,\ \frac{NX^2}{3}\right)$	X is the range of the original distribution
U-quadratic	$e_o \sim N\left(0,\ \frac{3NX^2}{5}\right)$	X is the range of the original distribution

The previous formalization can also be rewritten as follows:

$$\sum_{i=1}^{N} c_i = \sum_{i=1}^{N} (c_i + x_i) - e_o$$

where e_o is the obtained error by the addition of random numbers. Therefore, e_o is the sum of all added random values:

$$e_o = \sum_{i=1}^{N} x_i.$$

Many analytical models using probability theory for different distributions were developed, as presented in Table 2. There are studies in the literature proposing optimization in the noise generation for some contexts [30]. However, in our study [4] for our context, we concluded that there are no differences for privacy and utility between these probability distributions. Therefore, we will demonstrate the masking approach using the Laplace distribution but the same procedure is also valid for other probability distributions.

Let x_i be a random variable generated from a Laplace distribution. Its variance is $\sigma_x^2 = 2b^2$, where b is a scale parameter. Now, for a large N, the central limit theorem ensures that the obtained error for billing purpose follows a normal distribution with mean $\mu_{e_o} = 0$ and variance:

$$\sigma_{e_o}^2 = N^2(\sigma_x^2/N) = N\sigma_x^2 = 2Nb^2. \tag{1}$$

In other words, to have an obtained error between two accepted values (with high probability), we can use the following normal distribution:

$$e_o \sim N\left(0,\ 2\,Nb^2\right).$$

As an example, suppose that the power provider wants to compute the total consumption of a consumer at the end of a month with 31 days. With measurements of 10 min, it has a total of $N = 4,464$ measurements. If the maximum allowed error e_a is 2 kWh, we have to find the variance $\sigma_{e_o}^2$ of the normal distribution such that the probability of the obtained error lies between -2 and 2 kWh is high, *e.g.*, 0.98 ($Pr(-2 \le e_o \le 2) = 0.98$). Using the cumulative density function of the normal distribution, this variance is $\sigma_{e_o}^2 = 0.739113$. Therefore, the meter would add a Laplacian noise with magnitude of $b = \sqrt{\sigma_{e_o}^2/(2N)} = \sqrt{0.739113/(2 \cdot 4,464)} = 0.0091$ to obfuscate the real consumption. In Section 5 we demonstrate how this parameter affects the achieved privacy level.

A key feature of the approach is the possibility to allow the consumer and the power provider to negotiate the privacy and utility levels. Moreover, since this masking approach considers only the data that is disclosed to the power provider, it does not affect the customer's ability of analyzing his own real consumption profile inside his home and identifying appliance usage [33].

Since to preserve privacy the proposed approach just generates a random number, we claim that the proposed approach is lightweight. In research of security and privacy, the computational complexity to generate a random number is considered as $O(1)$. Therefore, the computational complexity of our proposed approach is $O(1)$.

5. Privacy validation

5.1. Differential privacy

Dwork [12] proposed the notion of differential privacy for general datasets. A mechanism of obfuscation is differentially private if its outcome is not significantly affected by the removal or addition of a single dataset participant. Here, we instantiate this notion for datasets of energy consumption profiles. Thus, appliances are participants in consumption profiles and an adversary learns approximately the same information about any individual appliance, regardless of its presence or absence in the original profile.

Considering that a consumption profile is a set of appliances, we say profiles $P1$ and $P2$ differ in at most one element if one is a proper subset of the other and the larger dataset profile contains just one additional appliance.

Definition 1. An obfuscation mechanism K gives ϵ-differential privacy if for all profiles $P1$ and $P2$ differing in at most one element, and all $S \subseteq Range(K)$,

$$Pr[K(P1) \in S] \leq exp(\epsilon) \times Pr[K(P2) \in S]$$

where the probability is taken over the randomness of K.

The ϵ value is the privacy metric and for better privacy, a small value is desirable. A mechanism K satisfying this definition addresses concerns that any appliance might have about the leakage of its information: even if the appliance has been removed from the dataset, no outputs (and thus consequences of outputs) would become significantly more or less likely.

Differential privacy is achieved by the addition of noise whose magnitude is a function of the largest change a single appliance could have on the output profile; this quantity is referred as the sensitivity of the function.

Definition 2. For f: $P \rightarrow R^k$, the sensitivity of f is

$$\Delta f = \max_{P1,P2} \parallel f(P1) - f(P2) \parallel_1$$

for all $P1$, $P2$ differing in at most one appliance.

In particular, when $k = 1$ the sensitivity of f is the maximum difference in the values that the function f may take on a pair of profiles that differ in only one appliance.

The privacy mechanism, denoted K for a query function f, computes $f(X)$ and adds noise with a Laplace distribution with mean $\mu = 0$ and scale parameter b:

$$b = \Delta f/\epsilon \ \ \therefore \ \ \epsilon = \Delta f/b. \tag{2}$$

The variance of the noise distribution is $2b^2$. On query function f the privacy mechanism K responds with

$$f(X) + (Lap(\Delta f/\epsilon))^k$$

adding noise with distribution $Lap(\Delta f/\epsilon)$ independently to each of the k components of $f(X)$. Note that decreasing ϵ, a publicly known parameter, flattens out the $Lap(\Delta f/\epsilon)$ curve, yielding larger expected noise magnitude.

Theorem 1. *For f: $P \rightarrow R^k$, the mechanism K that adds independently noise with distribution $Lap(\Delta f/\epsilon)$ gives ϵ-differential privacy.*

The proof of Theorem 1 is straightforward and Dwork gives this proof for general datasets in [13]. This theorem describes a relationship between Δf, b, and the differential privacy. To achieve ϵ-differential privacy, one must choose $b \geq \Delta f/\epsilon$. Given a sufficiently small ϵ, differential privacy limits the ability of an adversary to identify an appliance.

5.2. Privacy for appliances

We demonstrate the privacy and utility levels achieved with the approach presented in Section 4 through a billing example. For load monitoring and time-based tariff examples, see [6] and Section 8.

The data used in the next example are measurements collected at each one minute from a residential consumer.[1] Fig. 1a shows a daily profile of a residential consumer. There are 16 appliances in this consumption profile. However, the appliance with highest wattage is the laundry dryer.

Note that it is not hard for a user (or an automated mechanism) to identify the wattage of an appliance purchased. That said, the global sensitivity of the profile from Fig. 1a is the maximum variation of the appliance with highest wattage (laundry dryer):

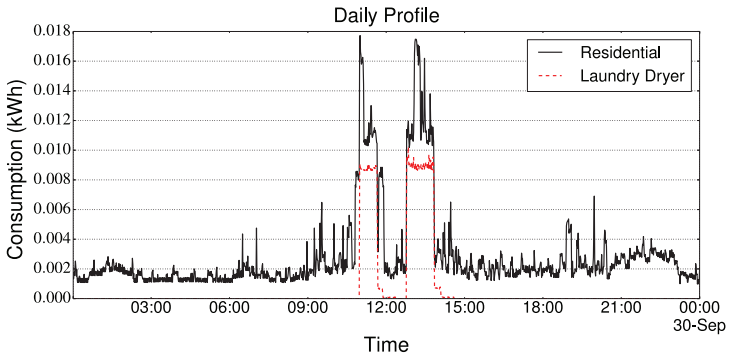
$$\Delta f = 0.01022.$$

In this example, we assume the billing period as one month. The total consumption of this consumer during the billing period (one month of 31 days) is 131.978 kWh. Considering a maximal allowed error of 5% for billing purpose, we have 6.5989 kWh. Thus, the variance for a high probability (*e.g.*, 0.98) of not exceeding this value is $\sigma^2_{e_o} = 8.04626$. Isolating the scale parameter b from Eq. 1, we have (for measurements at each 1 min, $N = 44,640$):

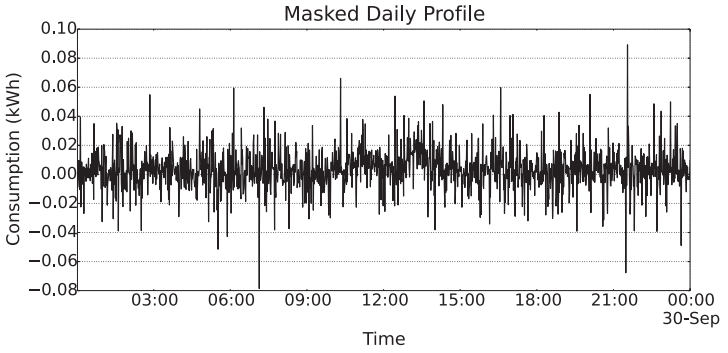
$$b = \sqrt{\sigma^2_{e_o}/(2\,N)} = 0.0094934.$$

Fig. 1b presents the daily masked profile using this Laplacian noise. Considering that at the end of the month the power provider sums the informed masked values by the consumer, it obtains a value of 133.7977 kWh. The real value is 131.978 kWh. The difference between these values is an error of 1.3788%, less than the maximum allowed error (5%).

[1] Combining appliance signatures we can generate arbitrary large populations and measurement frequency. Several databases of appliance signatures are available online (*e.g.* Tracebase [28]).



(a) Residential (black solid) and Laundry Dryer (red dashed) daily profiles with measurements at each 1 minute.



(b) Residential masked daily profile with measurements at each 1 minute.

Fig. 1. Effect of the noise addition.

This error may be less if the consumer does not mask the measurements all the time. Also, if the meter firmware accumulates the sum of the added random numbers and sends that with the last measurement of the month, the error is zero.

We can conclude that, from Eq. 2, using an utility requirement of 5%, the achieved privacy level in this example is:

$$\epsilon = \frac{\Delta f}{b} = 1.077.$$

In the literature of differential privacy, some researchers argue that the value of ϵ may be relative because for the same value of ϵ, the probability of identifying a participant is dependent on the context. Lee et al. [22] propose a technique to, in accordance with a context, find a suitable value of ϵ:

$$\epsilon = \frac{\Delta f}{\Delta v} \ln \frac{(n-1)p}{1-p} \tag{3}$$

where Δv is the largest distance of possible values, p is the probability of identifying the presence of an individual and n is the number of individuals in the data set. In the previous example for the consumption profile, we have $\Delta v = 0.0178$ (highest value presented in the profile) and $n = 16$ (number of appliances). Thus, considering that the probability of identifying the usage of an appliance is 1/3, it is suggested to have:

$$\epsilon \leq \frac{0.01022}{0.0178} \ln \frac{15 \cdot 3}{3 \cdot 2} = 1.1568.$$

Since in our example the obtained value of ϵ is 1.077, we can conclude that an attacker has a probability of identifying the laundry dryer less than 1/3. In other words, on average, he/she has to try more than three moments to be able to guess one moment of usage of the laundry dryer (and maybe he/she does not know that).

Rearranging Eq. 3, we can determine the probability:

$$p = \frac{1}{1 + (n-1) \cdot exp(-\frac{\epsilon \Delta v}{\Delta f})}. \tag{4}$$

In Table 3 we present the wattages and achieved privacy levels for some appliances of the profile from the previous example. The wattages were obtained from the Tracebase dataset [28], the ϵ values were obtained from Eq. 2, and the probability values of identifying appliances were obtained from Eq. 4.

Table 3
Privacy levels achieved for each appliance from the profile of Fig. 1a. Lower values of ϵ and p imply better privacy.

Appliance	Max. Wattage	Privacy (ϵ)	Probability (p)
Charger smartphone	~6	0.001	0.0626
Router	~9	0.003	0.0628
Playstation 3	~130	0.037	0.0663
TV LCD	~140	0.075	0.0706
PC desktop	~300	0.077	0.0708
Refrigerator	~1000	0.099	0.0733
Printer	~600	0.127	0.0767
Water fountain	~260	0.143	0.0787
Toaster	~700	0.257	0.0944
Cooking stove	~900	0.276	0.0973
Vacuum cleaner	~1100	0.336	0.1068
Iron	~1500	0.347	0.1087
Coffee maker	~1300	0.353	0.1097
Microwave oven	~1400	0.457	0.1287
Washing machine	~3000	0.903	0.2431
Laundry dryer	~3500	1.077	0.3031

As discussed before, lower values of ϵ or p imply better privacy. Naturally, the consumer behavior and privacy are correlated with the appliance usage and some of these information may be more sensitive than others.

As we can see in Table 3, the appliance wattage is very correlated with the privacy level. However, some appliances with higher wattage have better privacy than some appliances with lower wattage. This is because the accumulated consumption (kWh) is not only dependent on the wattage, but also on the time of use of the appliance.

It is known that less measurements implies higher privacy. Taking an extreme example, if a consumer sends to the power provider only one measurement with the total consumption at the end of the billing period, the achieved privacy is much better than sending measurements at each 1 min. However, when the number of measurements in a time period is large, our approach generates noise to hide each individual measurement, and the achieved privacy could be as higher as sending only one measurement with the total consumption at the end of the billing period.

Using the proposed model, it is also possible to calculate the achieved utility level for a given privacy requirement. From Eqs. (1) and (2) we have:

$$\sigma_{e_o}^2 = \frac{2 \cdot N \cdot \Delta f^2}{\epsilon^2}. \tag{5}$$

Therefore, with this variation, the maximum obtained error (utility level) is easily calculated using the inverse cumulative density function (quantiles) of the normal distribution.

6. Utility validation

To validate the utility, many features or benefits that use metering data were listed and evaluated to check whether they are still supported when using masked data. Since the masking approach preserves the aggregated values, the procedure to check if a feature is supported can be mapped as a checking if the feature uses only aggregated values or also uses individual values. For example, in energy theft/losses detection, Anas et al. [2] stipulate that losses consist of technical (*TL*) and non technical losses (*NTL*). Thus, we have:

$$Total\ Energy\ Losses = Energy\ Supplied - Bills\ Paid, \tag{6}$$

and also:

$$Total\ Energy\ Losses = NTL + TL. \tag{7}$$

Combining Eqs. 6 and 7 we get:

$$NTL = Energy\ Supplied - Bills\ Paid - TL. \tag{8}$$

Therefore, since all these variables consist of aggregated values, the feature of energy theft/losses detection can be regarded as supported.

Table 4 presents the list of features. The check if a feature uses only aggregated values (and thus if it is supported) can be found in the references. From the features that are not supported, “load forecasting for individual consumers” and “individual data analytics” can be considered as privacy threats that were solved. Since the approach masks the profile using noise addition, the individual demand or consumption levels are not original and the feature “demand-based rates” is not supported. Therefore, if the power provider wants to use this feature, it is a limitation.

Table 4
Metering data features and the masking impact.

Feature / Benefit	Support
Billing optimization [6]	Yes
Load monitoring and management for specific groups or regions [6]	Yes
Energy theft/losses detection [2]	Yes
Load forecasting for specific groups or regions [17]	Yes
Load forecasting for individual consumers [17]	No
Time-based rates (*e.g.*, different prices based on time of day and season) [6]	Yes
Demand-based rates (*e.g.* different prices based on demand levels) [27]	No
Individual data analytics (*e.g.* NIALM and marketers) [5,26]	No
In-home feedback tools: estimated bills, device profiles etc.[33]	Yes

Table 5
Filtering effect using different values of P. In this example, $P = 30$ is the best value.

P	Correlation	P	Correlation
0	0.2135	29	0.7363
2	0.4154	30	0.7368
4	0.5101	31	0.7362
8	0.6149	32	0.7351
16	0.6859	256	0.4164

Some other features that are not related with consumption measurement data, but are provided by the deployment of smart meters, are:

- Outage notifications;
- Remote disconnect/connect;
- Diagnosis of power quality problems;
- Reduction of costs of metering readings;
- Two way flow (possibility to consumers sell their produced electricity energy).

These features are still supported despite the masking.

7. Filtering attack

In previous work we evaluated two types of privacy attacks to the proposed solution. The first attack [6] was based on the hypothesis that a consumer tends to have a similar weekly behavior. Therefore the attacker may collect the consumer's data and calculate an expected week for this consumer. According to experiments this attack is unsuccessful because, even choosing a consumer who repeats a behavior almost always, it is better to guess the behavior using the own masked week than using the expected week.

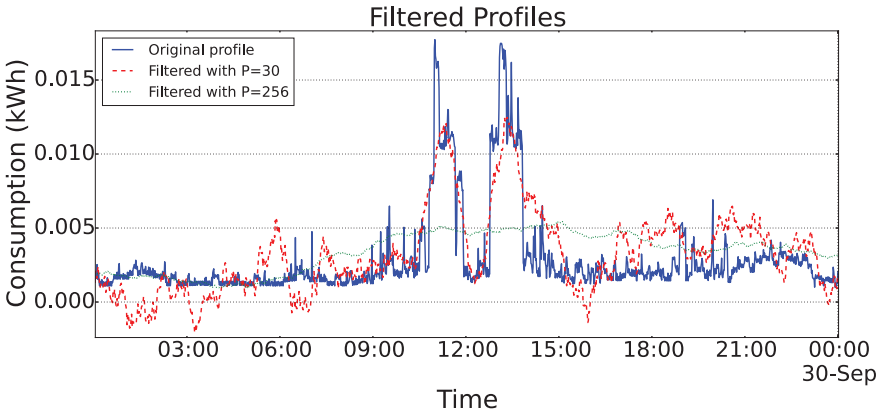
The second evaluated attack [5] was a Non-Intrusive Appliance Load Monitoring (NIALM). According to experiments, even using a high utility level and a low obfuscation (choosing a small allowed error e_a), the masking approach still inhibits the NIALM potential to detect appliance usages.

In this paper we present a filtering attack. It is based on the calculation of moving arithmetic means throughout the profile. The algorithm for the filtering attack works as follows:

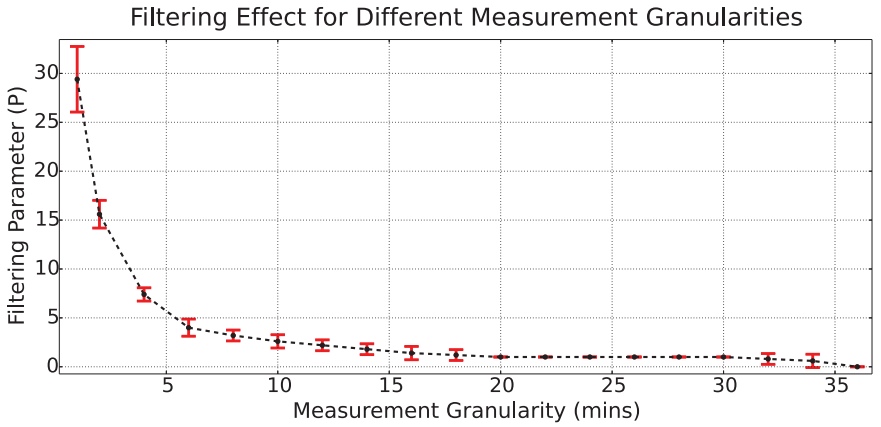
- Let T be a time series, which represents the masked profile and T_f a new series (the resulting filtered profile).
- The first P values from T_f will be equal to the first P values of T and the last P values from T_f will be equal to the last P values of T.
- The value with index $P+1$ from T_f will be the mean of the values with indexes from 1 to $2P+1$ from T. The value with index $P+2$ will be the mean of the values with indexes from 2 to $2P+2$ from T, and so on for all remaining values. This procedure creates moveable means to eliminate the high-frequency noise.

To analyze the attack effectiveness, we made experiments to verify if the correlation coefficient between the original profile and the masked profile is increased after the filtering. For example, using the masked profile from Fig. 1b, we verified if the added noise is removed and if the correlation with the original profile is increased. As presented in Table 5, we made this procedure using different P values. In this example, the configuration which presented best result was $P = 30$ and the filtered profile is presented by the dashed red line in Fig. 2a.

As observed in Table 5, the effectiveness of filtering is increased when we increment the P value because the high frequency noise is increasingly eliminated. In fact, the correlation between the masked profile and the original profile is



(a) Profile from Fig. 1b filtered with $P = 30$ (best filtering in this example) and $P = 256$.



(b) Filtering parameter (P) for different measurement granularities. The confidence intervals are of 95%.

Fig. 2. Filtering effect.

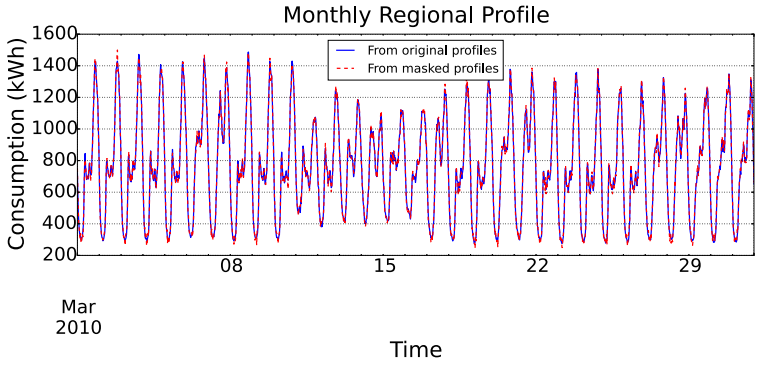
0.2135, whereas filtering with $P = 30$ we obtain a correlation of 0.7368. In other words, we can consider that this attack has some effect on privacy. However, the attacker may not know which P value to choose and if the chosen one is greater than 30, the obtained correlation is dwindled due the filtering saturation. In this example, choosing P values greater than 30, the filter will eliminate not only the added noise, but also the original characteristics of the profile, as can be observed by the dotted green line in Fig. 2a (filtered with $P = 256$).

Experiments also showed that coarse-grained measurements (with longer time intervals) implies in less filtering efficiency. This is because each measurement becomes less correlated with the adjacent measurements. For the same consumer of the previous example, the Fig. 2b presents the mean of the best values of P (vertical axis) for different granularity of measurements (horizontal axis). As observed for this consumer, using measurement intervals of 36 min (or greater) will make filtering attacks ineffective because the best setting is using a P equals to 0 (*i.e*, no filtering). This means that with a granularity greater than or equals to 36 min, the original profile is more correlated with the masked profile than with the filtered profile.

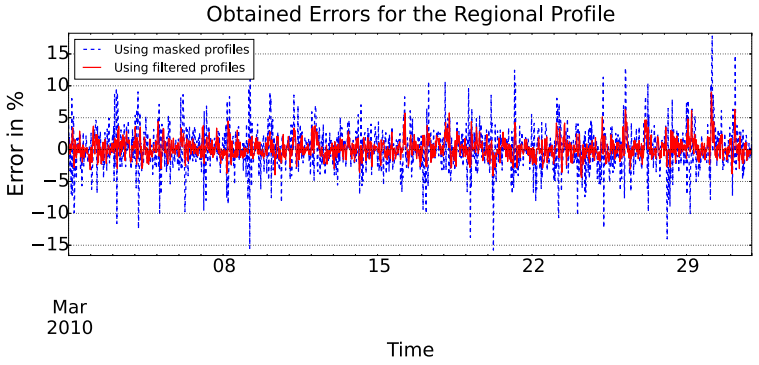
8. Utility improvement

If each consumer masks his data based on the billing period, the power provider may obtain accurate values for billing. However, to obtain accurate values for load monitoring in a region, the number of consumers should be as many as possible because it is desired that the added noise should be unnoticed in the aggregated data used for load monitoring.

The data used in the next example are measurements collected at each 30 min from real residential consumers (anonymised) from Ireland (CER) [11]. Suppose that the power provider wants to compute the total consumption in a region with many consumers through time for load monitoring (*e.g.*, find peak times, leak detection, load forecasting and many other applications). Using a billing period of 1 month and measurements at each 30 min, it has $N = 1488$ measurements for



(a) Regional profile using original data (blue solid) versus using masked data (red dashed) with measurements of 30 min. The profiles are very similar.



(b) Obtained errors for the regional profile using masked profiles versus using filtered profiles. Using filtered profiles implies in better accuracy for load monitoring.

Fig. 3. Regional consumption profile.

each consumer during a month with 31 days (March). So, in this experiment, we consider a region with 1488 consumers (thus, the data forms a square matrix). Fig. 3a shows the regional profile during March obtained from original consumer profiles *versus* the regional profile obtained from masked profiles.

As depicted in Fig. 3a, visually there is no difference between the two profiles. However, the obtained errors depend on the population behavior. For example, in a high consumption period the obtained error has a different proportion from the obtained error in a low consumption period. The large errors in Fig. 3a were obtained in periods of low consumption (*e.g.*, during the night), because while consuming less, consumers are still masking their data using a noise level based on the billing period. The blue dashed line in Fig. 3b presents the obtained errors through time for this scenario (Fig. 3a). These errors may be lower if not all consumers decide to mask their data all the time.

The possibility of having consumers deciding when to send masked (*vs.* original) measurements is in accordance with the privacy definition mentioned by Stallings et al. [31], which says: "*Privacy assures that individuals control or influence what information related to them may be collected and stored and by whom and to whom that information may be disclosed*". Since not all consumers are masking the measurements all the time (making the obtained error to be lower), this implies in a utility improvement.

Another utility improvement is instead of sending just masked profiles, sending filtered masked profiles. Considering that the attacker has the ability to discover and use the best filter setting and that even analyzing the filtered data, the privacy of the consumer is still preserved due the differential privacy constraint [25], one may argue that consumers may disclose already filtered profiles for utility improvement.

For other types of data sets (*i.e*, not electrical consumption data), Mivule et al. [25] suggested that, filtering differentially private data maintains some statistical properties such as sum. In fact, according to our experiments, there is no utility improvement for billing purposes because the computed total consumption in the end of the month using a filtered masked profile is statistically the same of using a non-filtered one. However, for regional load monitoring, which computes statistics using measurements of different profiles in an instant of time (*i.e*, not using measurements through time of only an auto-correlated profile as is the case of billing), the improvement of utility is significant.

Disclosing filtered profiles makes the obtained errors for the regional profile to be much lower, as presented by the solid red line in Fig. 3b. From 1488 monthly profiles of the CER dataset, the filtering mechanism was effective for only 386 of

them. Therefore, 338 of them increased the correlation filtering with $P = 1$, 34 with $P = 2$, 10 with $P = 3$ and only 4 with $P = 4$. For the remaining profiles the filtering was ineffective so it was considered that the consumers disclosed the masked profile instead of a filtered version.

9. Performance analysis

In Section 2 we presented the related work and defined the aspects of performance to compare different approaches to preserve privacy in smart metering. Since each approach has its advantages and disadvantages, we do not consider a specific one as the baseline for comparisons.

As presented in Table 1, the discussed aspects were computational complexity, scalability, meters' independence, cost, environmental impact and accuracy. Regarding these attributes, we make the following notes:

- Cost and environmental impact: as evidenced, we can assume that rechargeable batteries may be not successful on these attributes.
- Meters' independence and scalability: by construction, proposed homomorphic encryption protocols need communication between meters and/or many message exchanges between meters and the power provider. Therefore, these attributes are compromised when using homomorphic encryption approaches.
- Accuracy: As demonstrated, different from other approaches, our noise addition approach is not one hundred percent accurate (although the error may be controlled).
- Computational complexity: besides estimation through asymptotic complexity, we claim that experimental analysis is necessary to present concrete results when comparing different proposals.

In this section, we present an experiment with the goal to compare the response time (which is also related to computational complexity) of our approach with other existing approaches. We address the following research question:

- RQ_1: How is our approach in terms of response time when compared with others existing approaches and in different configuration scenarios?

We implemented simulators[2] in C programming language and executed in a machine with 1.6 GHz Intel Core i5 processor and 6 GB of RAM memory and Ubuntu 14.04 operational system.

In total we implemented 5 approaches of privacy preserving in smart metering. One using noise addition (our approach), other using rechargeable batteries and the three others using homomorphic encryption.[3] All the simulations were considering load monitoring in a region with N meters.

The implemented approach that uses rechargeable batteries was the one proposed by McLaughlin et al. [24], called Non-Intrusive Load Leveling (NILL). The goal of a NILL system is to level the load profile to a constant *target load*, thus removing appliance signatures. When an appliance turns ON, it will exert a load beyond the target load. Thus, NILL will discharge the battery to partially supply the load created by the appliance, maintaining the target load. Similarly, if an appliance enters the OFF state, the load profile will decrease below the target load. These opportunities are used to charge the battery while restoring the target load.

The implemented homomorphic encryption approaches were:

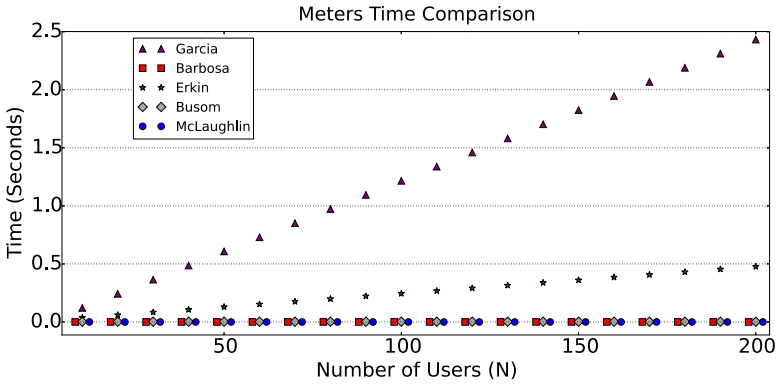
- Garcia et al. [15]: Combines the Paillier additive homomorphic encryption and additive secret sharing. The proposal completely hides the measurements from the power provider, since it receives encrypted measurements that it cannot decrypt, and random shares of the total consumption. At the same time, neither of the participants can retrieve meaningful information on the consumption of others as they only see the random shares.
- Erkin et al. [14]: Combines a modified version of the Paillier encryption with hashes generated from synchronized timestamps of all meters. Smart meters communicate with each other but no one can decrypt the encryptions. However, an aggregator (the power provider or anyone in the group of meters) can collect the encryptions and form the total consumption.
- Busom et al. [10]: Uses a modified version of the ElGamal encryption and an algorithm to brute force the discrete logarithm for a known range of values. It does not require communication between the meters but still needs many message exchanges between meters and the power provider.

In our simulations, using different configuration scenarios (number of meters, from 1 to 200) to calculate the total consumption in the region, we measured the processing time of each meter (Fig. 4a) and the aggregator (Fig. 4b). From the evaluation results, we make the following observations:

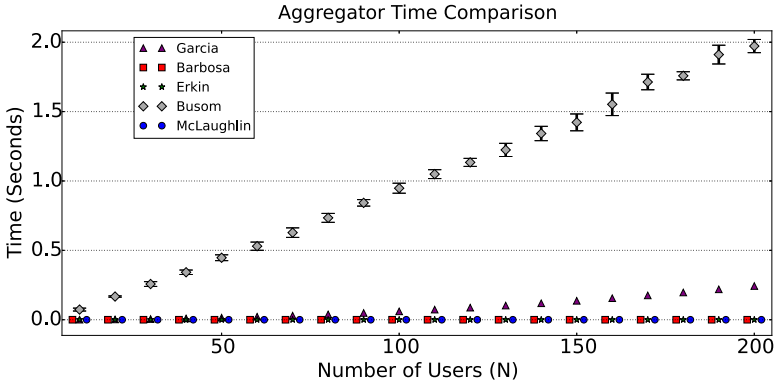
- RQ_1: How is our approach in terms of response time when compared with others existing approaches and in different configuration scenarios?

[2] The source codes can be found at our GitLab repository (https://git.lsd.ufcg.edu.br/pedroysb/privacy-performance-smart-metering).

[3] To implement these simulators, we used functions of the *libgmp* (https://gmplib.org), *libpaillier* (http://acsc.cs.utexas.edu/libpaillier), and *libcrypto* (https://www.openssl.org/docs/manmaster/crypto/crypto.html).



(a) Processing time of smart meters in 5 different privacy approaches.



(b) Processing time of aggregator in 5 different privacy approaches.

Fig. 4. Processing time analysis.

– Our noise addition approach and the one that uses rechargeable batteries presented very low response times, whereas the homomorphic encryption approaches presented considerable delays.

It is also important to note that there are many message exchanges in the homomorphic encryption approaches and we are not considering possible network delays in our experiments.

Each configuration was executed 10 times and the average values are being considered. These amounts of repetitions were enough to get precise average values and due to the very low obtained variations, the confidence intervals are being omitted here, except for the aggregation in the approach proposed by Busom et al. [10], which needs a brute forcing to solve the discrete logarithm problem. The confidence intervals in these cases are of 95%.

Despite the response time of the rechargeable batteries approach be as good as our approach, it is hard to ignore the environmental effects and the costs of using batteries. Therefore, we consider that, in many cases, noise addition stands in relation to other approaches.

10. Conclusions

A lightweight and efficient privacy-preserving metering scheme for Smart Grids was proposed and evaluated. We claim that the solution meets the needs of consumers (privacy) and power providers (utility). The modification in the communication procedure between a smart meter and the power provider is just the generation of a random number and the addition of this number to the measurement to be sent to the power provider. Therefore, due the simplicity and low complexity, the approach can be deployed in low-cost microcontrollers.

Using the scheme, the differential privacy guarantee for appliances is achieved. After evaluating the experimental results with consumer profiles, common Smart Grid applications, and after comparing with other existing approaches, we conclude that our solution can be regarded as useful.

Furthermore we also presented an attack to the solution based on filtering. Given a masked profile of a consumer, the attacker may try to filter the random noise using an algorithm of moveable means. This attack has some effect on the privacy level, but also can be used for good purposes. For example, if consumers disclose already filtered profiles or if the power provider filters that when receiving, the obtained errors for load monitoring in a region may be lower.

Socially speaking, it is important to note that all presented experiments and examples were considering the worst case: users always disclosing masked data. However, with the privacy solution in a real product, some users may want to enjoy the Smart Grid benefits and do not mask or mask the data only in private and sensitive moments. Therefore, the masking could be enabled or disabled through a checkbox in a graphical interface or through a physical switch button in the meter. A regulatory agency could certify that the button and the masking are working properly.

Allowing the enabling or disabling of masking would improve the utility in both dimensions: for billing, since a consumer would not mask all the time, and for load monitoring in a region, since not all consumers would mask in an instant of time. We let this social and usability study as a future work. We also let the implementation of the approach in real smart metering products and the study of more possible attacks as future works.

References

[1] G. Ács, C. Claude, I have a dream! (differentially private smart metering), in: Proc. of the 13th International Conf. on Information Hiding, 2011, pp. 118–132. Prague, Czech Republic. doi: 10.1007/978-3-642-24178-9_9.

[2] M. Anas, N. Javaid, A. Mahmood, S.M. Raza, U. Qasim, Z.A. Khan, Minimizing electricity theft using smart meters in ami, in: Proc. of the IEEE Seventh International Conf. on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2012, pp. 176–182. Victoria, Canada. doi: 10.1109/3PGCIC.2012.42.

[3] M. Backes, S. Meiser, Differentially private smart metering with battery recharging, in: Data Privacy Management and Autonomous Spontaneous Security, 2014, pp. 194–212, doi:10.1007/978-3-642-54568-9_13.

[4] P. Barbosa, Preservando a Privacidade em Smart Grids Através de Adição de Ruído, Master's thesis Federal University of Campina Grande, Campina Grande, Brazil, 2014.

[5] P. Barbosa, A. Brito, H. Almeida, Defending against load monitoring in smart metering data through noise addition, in: Proc. of the 30th Annual ACM Symposium on Applied Computing (SAC), 2015, pp. 2218–2224. Salamanca, Spain. doi: 10.1145/2695664.2695800.

[6] P. Barbosa, A. Brito, H. Almeida, S. Clauß, Lightweight privacy for smart metering data by adding noise, in: Proc. of the 29th Annual ACM Symposium on Applied Computing (SAC), 2014, pp. 531–538. Gyeongju, South Korea. doi: 10.1145/2554850.2554982.

[7] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava, Nilmtk: an open source toolkit for non-intrusive load monitoring, in: 5th International Conf. on Future Energy Systems (ACM e-Energy), 2014. Cambridge, UK. doi: 10.1145/2602044.2602051.

[8] C. Boccuzzi, Smart grid e o big brother energético, Metering International América Latina 3 (2010) 82–83.

[9] J. Bohli, C. Sorge, O. Ugus, A privacy model for smart metering, in: Proc. IEEE International Conf. Communications Workshops (ICC), 2010, pp. 1–5. Cape Town, South Africa. doi: 10.1109/ICCW.2010.5503916.

[10] N. Busom, R. Petrlic, F. Sebé, C. Sorge, M. Valls, Efficient smart metering based on homomorphic encryption, Comput. Commun. (2015) 95–101, doi:10.1016/j.comcom.2015.08.016.

[11] CER, Commission for energy regulation – cer smart metering project. smart meter electricity trial data, 2012.

[12] C. Dwork, Differential privacy, in: Proc. of the 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP), 2006, pp. 1–12. Venice, Italy. doi: 10.1007/11787006_1.

[13] C. Dwork, Differential privacy: a survey of results, in: Proc. of the 5th International Conf. of Theory and Applications of Models of Computation (TAMC), 2008, pp. 1–19. Xi'an, China. doi: 10.1007/978-3-540-79228-4_1.

[14] Z. Erkin, G. Tsudik, Private computation of spatial and temporal power consumption with smart meters, in: Proc. of the 10th Int. Conf. on Applied Cryptography and Network Security (ACNS), 2012, pp. 561–577, doi:10.1007/978-3-642-31284-7_33.

[15] F.D. Garcia, B. Jacobs, Privacy-friendly energy-metering via homomorphic encryption, in: Security and Trust Management, 6710, 2010, pp. 226–238, doi:10.1007/978-3-642-22444-7_15.

[16] X. He, X. Zhang, C.C.J. Kuo, A distortion-based approach to privacy-preserving metering in smart grids, IEEE Access 1 (2013) 67–78, doi:10.1109/ACCESS.2013.2260815.

[17] D. Ilić, P.G. Silva, S. Karnouskos, M. Jacobi, Impact assessment of smart meter grouping on the accuracy of forecasting algorithms, in: Proc. of the 28th Annual ACM Symposium on Applied Computing (SAC), 2013, pp. 673–679. Coimbra, Portugal. doi: 10.1145/2480362.2480491.

[18] M. Jelasity, K.P. Birman, Distributional differential privacy for large-scale smart metering, in: Proc. of the 2nd ACM workshop on Information hiding and multimedia security, 2014, pp. 141–146, doi:10.1145/2600918.2600919.

[19] G. Kalogridis, C. Efthymiou, S.Z. Denic, T.A. Lewis, R. Cepeda, Privacy for smart meters: towards undetectable appliance load signatures, in: IEEE 1st International Conf. on Smart Grid Communications (SmartGridComm), 2010, pp. 232–237. Gaithersburg, USA. doi: 10.1109/SMARTGRID.2010.5622047.

[20] O. Koehle, Just Say No to Big Brother's Smart Meters. The Latest in Bio-Hazard Technology, ARC Reproductions, 2012.

[21] K. Lauter, M. Naehrig, V. Vaikuntanathan, Can homomorphic encryption be practical? in: Proc. of 3rd ACM workshop on Cloud computing security (CCSW), 2011, pp. 113–124. Illinois, USA. doi: 10.1145/2046660.2046682.

[22] J. Lee, C. Clifton, How much is enough? choosing ϵ for differential privacy, in: Proc. of the 14th International Conf. on Information Security, in: ISC'11, 2011, pp. 325–340. Xi'an, China. doi: 10.1007/978-3-642-24861-0_22.

[23] X. Liao, D. Formby, C. Day, R.A. Beyah, Towards secure metering data analysis via distributed differential privacy, in: Proc. of the 44th IEEE/IFIP International Conference on Dependable Systems and Networks, 2014, pp. 780–785, doi:10.1109/DSN.2014.82.

[24] S. McLaughlin, P. McDaniel, W. Aiello, Protecting consumer privacy from electric load monitoring, in: Proc. of the 18th ACM Conf. on Computer and Communications Security (CCS), 2011, pp. 87–98. Illinois, USA. doi: 10.1145/2046707.2046720.

[25] K. Mivule, C. Turner, Applying moving average filtering for non-interactive differential privacy settings, in: Procedia Computer Science, 36, 2014, pp. 409–415, doi:10.1016/j.procs.2014.09.013.

[26] NIST, National institute of standards and technology – guidelines for smart grid cybersecurity: Vol. 2 - privacy and the smart grid, 2014.

[27] Procel, Manual de tarifação da energia elétrica. programa nacional de conservação de energia, Eletrobras, 2011.

[28] A. Reinhardt, P. Baumann, D. Burgstahler, M. Hollick, H. Chonov, M. Werner, R. Steinmetz, On the accuracy of appliance identification based on distributed load metering data, in: Proc. of the 2nd IFIP Conf. on Sust. Internet and ICT for Sustainability, 2012, pp. 1–9.

[29] L. Sankar, S.R. Rajagopalan, S. Mohajer, H.V. Poor, Smart meter privacy: a theoretical framework, IEEE Trans. on Smart Grid 4 (2013) 837–846, doi:10.1109/TSG.2012.2211046.

[30] J. Soria-Comas, J. Domingo-Ferrer, Optimal data-independent noise for differential privacy, Inf. Sci. 250 (2013) 200–214, doi:10.1016/j.ins.2013.07.004.

[31] W. Stallings, L. Brawn, Computer Security: Principles and Practice, 3, Pearson, 2015.

[32] S. Wang, L. Cui, J. Que, D.-H. Choi, X. Jiang, L. Xie, A randomized response model for privacy preserving smart metering, IEEE Trans. on Smart Grid 3 (2012) 1317–1324, doi:10.1109/TSG.2012.2192487.

[33] L. Ying-Xun, L. Chin-Feng, H. Yueh-Min, C. Han-Chieh, Multi-appliance recognition system with hybrid svm/gmm classifier in ubiquitous smart home, Inf. Sci. 230 (2013) 39–55, doi:10.1016/j.ins.2012.10.002.

[34] J. Zhao, T. Jung, Y. Wang, X. Li, Achieving differential privacy of data disclosure in the smart grid, in: Proc. of IEEE INFOCOM, 2014, pp. 504–512, doi:10.1109/INFOCOM.2014.6847974.