CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 4
29-31 October 2019, the Hague, the Netherlands 16 September 2019

Protecting Consumer Privacy in Smart Metering by Randomized Response

Bastian Stölb, Josep Domingo-Ferrer and David Sánchez (Universitat Rovira i Virgili, Catalonia)
bastiannikolas.stolb@estudiants.urv.cat, josep.domingo@urv.cat, david.sanchez@urv.cat

Abstract and Paper

In recent years, an increased interest in Smart Meter (SM) privacy has emerged. In spite of their many advantages, like matching energy production with consumption, SMs pose serious concerns regarding the privacy of energy consumers. The vast majority of the work in this area has focused on cryptographic solutions. Due to resource constraints of SMs, cryptography-based methods are too heavy for them. Non-cryptographic approaches mostly rely on differential privacy (by adding noise to the real query result). Previous research in this area has been restricted to data aggregation. The aim of this paper is to propose a novel scheme to maintain privacy, based on randomized response. At the same time, this approach is appealing due to its simplicity, preservation of individual readings, efficient computation and communication. For a realistic scenario, real SM data from ESSnet (European Statistical System) Big Data is used.

Keywords

Smart Meter, Smart Grid, Privacy, Randomized Response, Consumer privacy, Data protection, Privacy Enhancing Technologies (PET)

Protecting Consumer Privacy in Smart Metering by Randomized Response

Bastian Stölb and Josep Domingo-Ferrer

Universitat Rovira i Virgili,
Department of Computer Science and Mathematics,
UNESCO Chair in Data Privacy,
Av. Països Catalans 26, 43007 Tarragona, Catalonia,
bastian.stoelb@estudiants.urv.cat, josep.domingo@urv.cat

Abstract. In recent years, an increased interest in smart meter (SM) privacy has emerged. In spite of their many advantages, like matching energy production with consumption, SMs pose serious concerns regarding the privacy of energy consumers. The vast majority of the work in this area has focused on cryptographic solutions. Due to resource constraints of SMs, cryptographic methods are often too heavy for them. Non-cryptographic approaches mostly rely on differential privacy (by adding noise). No matter whether cryptography is used or not, state-of-the-art research tends to be restricted to offering privacy-protected aggregations of SM readings. The aim of this paper is to propose a novel scheme to maintain privacy of SM readings based on Randomized Response (RR). The proposed RR approach is appealing due to its simplicity, its ability to preserve the distribution of individual readings, and its efficient computation and communication. Experimental results on real SM data from the London Data Store are presented.

1 Introduction

By the end of 2018, roughly 44 percent of European households were equipped with a smart meter (SM). According to [13], by the year 2023 the distribution will be around 71 percent, which means that over 200 million homes in Europe will be equipped with smart meters.

With SMs the power usage can be observed in real time by the energy providers, thus allowing as much electricity as necessary to be produced in a specific time. However, this increase of efficiency comes at the cost of privacy. Specifically, the number of people in an accommodation at a given moment

can be found out by using smart meter data. This invites burglars in case the power consumption of a flat is known to be very little or nothing.

Also, the habits of the residents can lead to serious privacy concerns: as stated by Garcia and Jacobs in [11] a person who wakes up at five o'clock in the morning in combination with a foreign name could possibly be identified as a religious Muslim. At a lower level, a so-called Nonintrusive Appliance Load Monitoring (NALM) [12] can even observe the electricity usage in detail. By using signatures of electronic devices (iron, refrigerator, dishwasher, microwave, hair dryer, etc.) it is possible to determine which device is turned on or off, with notable precision.

Other privacy risks and numerous conflicts of interests are outlined by Anderson and Fuloria in [2]. Among others, the researchers give the following suggestions: "smart meter data should belong to the customer[1]". The electricity provider needs access to these data only for supply and accounting reasons. A common database is rejected by the authors; instead they demand "a framework of standards that allow data to be shared between energy suppliers, distributors and management companies[2]".

2 Non-Cryptographic Smart Meter Privacy

The communication of smart meters with utility companies takes place over a public network (the Internet) and is regarded as unsafe unless appropriate countermeasures are taken. That is why a lot of work in this area has been performed and is still ongoing on protecting privacy. Most researchers focus on cryptographic solutions, mainly Partially Homomorphic Encryption (PHE) and Fully Homomorphic Encryption (FHE). Although cryptographic methods are appealing, one should consider the limited computational capacity of a smart meter. With this in mind, non-cryptographic solutions seem worth exploring.

2.1 Privacy models

Non-cryptographic solutions are normally oriented to satisfying a privacy model, that is, a privacy condition. As stated by Domingo-Ferrer and Soria-Comas in [8], there are four major privacy models:

- *Randomized Response.* Randomized Response (RR) was invented by Warner in 1965 [17] and views privacy as deniability. RR allows obtaining answers to sensitive questions (e.g. did you take drugs?) while

[1] [2], p. 16
[2] [2], p. 16

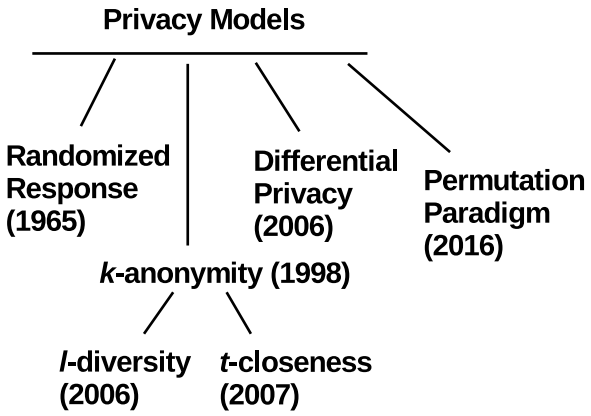

Figure 1: Privacy models

assuring the respondent's privacy. The respondent is allowed to randomize her answer according to some prescribed probability matrix before reporting it. Thus, she can deny that the reported answer is her true answer, but at the same time the interviewer can use his knowledge of the probability matrix to estimate the true distribution of the answers of a set of respondents from the distribution of their reported answers.

- *k-anonymity.* k-anonymity [15] was designed for privacy-preserving release of data sets. Assume the attributes in a data set can be split into identifiers (those that directly identify the respondent to whom a record corresponds), quasi-identifiers (attributes that do not uniquely identify the respondent when considered separately, but whose combination may, like Age, Zipcode and Job) and confidential attributes (those that carry confidential information, such as Diagnosis or Salary). A data set whose identifiers have been removed is said to be k-anonymous if any combination of values of quasi-identifier attributes in it is shared by at least k records. This provides protection against re-identification of the subjects to whom the records correspond. Bohli et al. [3] contend that the concept of k-anonymity is not suited to the smart metering problem, as there is no central entity releasing the smart meter readings.

- *Differential Privacy.* According to [9] the output of a query to a database is said to be differentially private if from its result it is not possible (up to a parameter ϵ) to notice the presence or absence of any specific record in the database. Usually differential privacy is achieved

by adding noise to the real query result.

- *Permutation Paradigm.* In [7], the authors presented the permutation paradigm, whereby they showed that any statistical disclosure control methods essentially consists of permutation plus perhaps a small noise addition.

2.2 Aggregation and smart meters

Aggregation is a useful principle to achieve privacy. It is also instrumental to reduce computation and communication costs, which is very welcome in smart meters because they have constrained resources. We summarize the overview on aggregation in smart meters given by Erkin and Tsudik in [10]:

- *Spatial aggregation.* In this approach smart meters are (geographically) clustered. The data of, say, a block of houses can be accumulated and presented to the energy supplier. This guarantees that individual households are protected, while still enabling load balancing.

- *Temporal aggregation.* In this case smart meters withhold data until a specific time interval has passed. Fine-grained data is summed up by the smart meter itself and protects the privacy of the customers. For the purpose of accounting, monthly aggregation is sufficient. However, many energy providers would like to have a consumption report at least on a daily basis.

- *Spatio-temporal aggregation.* In this hybrid setting, both approaches are combined. To monitor the consumption at specific times without encroaching on consumer privacy, the data of several SMs are spatially aggregated. For the purpose of billing, the measurements of an individual SM are aggregated over time.

3 Related Work

According to Bohli, Sorge and Ugus [3], privacy can be defined by using a "right-or-left type of [cryptographic] game[3]". Two proposals are offered: one of them includes a Trusted Third Party (TTP), while the other does not. In the first scenario the researchers suggest to use a TTP for aggregation purposes. The data are encrypted by the smart meter before it sends them to the TTP. The second is more frugal: every SM perturbs the measured

[3] [3], p. 2

values by adding random noise (with a specific distribution). In case of malfunction of a single SM, the energy provider is aware of missing data and the total usage can be roughly estimated. To achieve sufficient privacy, it must be assured that the added random values are considerably large, which is a drawback.

In [1] the authors suggest to use the differential privacy model. The proposal is straightforward, economic and easy to implement. There is no need for a Trusted Third Party. Individual smart meters add Laplacian noise to the measured data, before a stream cipher is applied. This results in quite low computational costs. The smart meters are clustered[4] and aggregated. The aggregator receives the accumulated values of all smart meters in the cluster. For the aggregator it is impossible to learn the consumption of an individual smart meter (at a specific time), which is good for privacy. However, the coarse resolution is disadvantageous. Also, this approach cannot cope with malfunctions. In case a single smart meter fails, the data of the whole cluster are lost, because of the stream cipher. Assuming the faulty smart meter is able to store the value and retransmit it later, the reliability of the proposed scheme can be improved.

A recent paper is [14]. The author presents an overview of ongoing research in the area of SM. To that end, he analyzes 53 papers (thereof 8 surveys) published in the last 9 years. He divides the studies into two main groups: attributable (with and without aggregation) and non-attributable (for accounting and operational purposes). He investigates whether a Trusted Third Party (TTP) is used. New approaches, like incentive-based or reward schemes, where users voluntarily share their data with their utility provider are emphasized. Although the paper provides a great overview about the state of the art in smart meter research, it does not distinguish between cryptographic and non-cryptographic solutions.

Finally, the following two papers claim to use randomized response for SM privacy. Wang et al. describe a model [16], in which a single SM can send the true data with a known probability. This approach requires so called Load Serving Entities (LSE). These LSEs can calculate the aggregated usage of a region, by using a statistical inference algorithm. A recent publication from 2018 by Cao et al. [4] suggests to add RR noise into the signature matrix of the behavior of the electronic devices. Furthermore the researchers recommend a technique to model the signature of behavior, by using sparse coding. Thus, a dictionary is generated after a short training period, containing the attributes of the electronic devices. However, these two papers use the term randomized response in the title only. They do not have anything

[4]An urban district consists of several hundreds or thousands of smart meters.

in common with the RR technique described by Warner.

4 Our Proposed Scheme

4.1 Goals

We propose a non-cryptographic solution that offers the following features:

- privacy;
- simplicity (lightweight, no need for a TTP);
- support for individual readings (without aggregation);
- low computational overhead;
- low communication overhead;
- integrity;
- high accuracy.

We demonstrate our approach with real data (from the London Data Store[5]) and using a realistic smart meter architecture (with respect to scalability).

4.2 Security model

We consider the following threats:

- eavesdropping;
- tampering;
- internal attacks.

However, the smart meters themselves will be assumed to be trusted devices.

4.3 Rationale

The diagram in Figure 2 compares cryptographic and non-cryptographic solutions. Our aim is to develop a scheme based on RR that can handle individual readings and maintain privacy at the same time.

[5] `https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households`

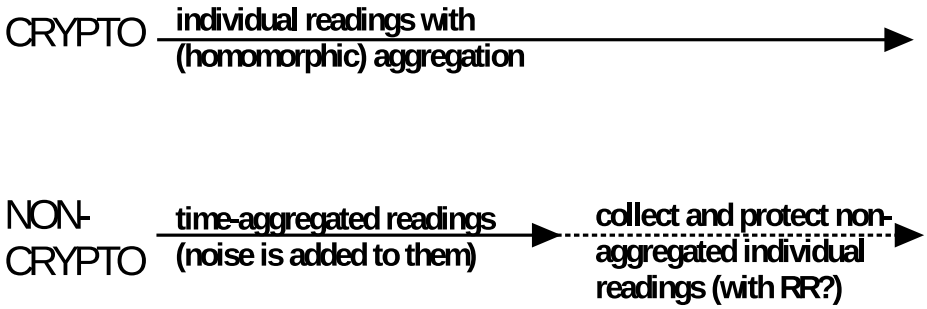

Figure 2: Cryptographic vs. non-cryptographic solutions

4.4 Background on randomized response

We briefly recalled randomized response in Section 2.1. Since we are going to use this technique in the rest of this paper, we now give more background on it (see [6] for further details).

Let us denote by X the attribute containing the answer to a sensitive question. If X can take r possible values, then the randomized response Y reported by the respondent instead of X follows an $r \times r$ matrix of probabilities

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \vdots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix} \tag{1}$$

where $p_{uv} = Pr(Y = v|X = u)$, for $u, v \in 1, \ldots, r$ denotes the probability that the randomized response is v when the respondent's true attribute value is u. Let $\pi_1, \ldots, \pi_r$ be the proportions of respondents whose true values fall in each of the r categories of X and let $\lambda_v = \sum_{u=1}^{r} p_{uv}\pi_u$ for $v = 1, \ldots, r$, be the probability of the reported value Y being v. If we define $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ and $\pi = (\pi_1, \ldots, \pi_r)^T$, it holds that $\lambda = \mathbf{P}^T\pi$. Furthermore, if $\hat{\lambda}$ is the vector of sample proportions corresponding to λ and $\mathbf{P}$ is nonsingular, in Chapter 3.3 of [5] it is proven that an unbiased estimator π can be computed as

$$\hat{\pi} = (\mathbf{P}^T)^{-}\hat{\lambda} \tag{2}$$

and they also provide an unbiased estimator of the dispersion matrix. In particular, the larger the off-diagonal probability mass in $\mathbf{P}$, the more dispersion (and the more respondent protection).

The commonalities and differences between RR and PRAM (Post RAndomization Method) are pointed out in [8]. Using PRAM the randomization

is applied by the data controller after collecting the true data (hence the term Post Randomization).

5 Experimental Results

5.1 Smart meter data set

As stated above, we have used real data from the London Data Store[6]. The UK Power Networks-led "Low Carbon London" project monitored from November 2011 to February 2014 the energy consumption of 5567 London households. The households were selected as a balanced sample of London's society. Data sets were then released with the energy consumption measured every half hour, a unique household identifier and a time-stamp. For this paper the data set UKPN-LCL-smartmeter-sample.csv was used. This file contains 17458 records of a single household. The highest measured value is 1.529 KWH/hh. The data have been cleansed by us – Null (representing a communication failure of the SM) values have been replaced by NaN (Not a Number). The first five entries can be seen in Table 1.

Table 1: Smart meter data from the London Data Store (extract)

	LCLid	DateTime	KWH/hh (per half hour)
1	MAC003718	17/10/2012 13:00:00	0.090
2	MAC003718	17/10/2012 13:30:00	0.160
3	MAC003718	17/10/2012 14:00:00	0.212
4	MAC003718	17/10/2012 14:30:00	0.145
5	MAC003718	17/10/2012 15:00:00	0.104

5.2 Creating the RR matrix

We have constructed RR probability matrices with $p = 0.4$, $p = 0.6$ and $p = 0.8$ in the main diagonal. We have filled the off-diagonal probabilities by using three different attenuation formulae A, B and C on p. In total we have 9 different probability matrices, corresponding to each value of p combined with each attenuation formula. Attenuations are as follows:

- *Attenuation formula A.* Populate the matrix with p in the main diagonal. For the super-diagonal and the sub-diagonal divide p by 2. For

[6] `https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households`

the super-super-diagonal and the sub-sub-diagonal divide p by 4, and so on. A 4×4 matrix would be as follows:

$$\mathbf{P} = \begin{pmatrix} p & p/2 & p/4 & p/8 \\ p/2 & p & p/2 & p/4 \\ p/4 & p/2 & p & p/2 \\ p/8 & p/4 & p/2 & p \end{pmatrix} \tag{3}$$

Finally, rescale the probabilities in each row of the matrix for them to sum to 1.

- *Attenuation formula B.* Populate the matrix with p in the main diagonal. For the super-diagonal and the sub-diagonal divide p by 1 plus the distance to the diagonal. Do the same for the super-super-diagonal and the sub-sub-diagonal, and so on. A 4×4 matrix would be as follows:

$$\mathbf{P} = \begin{pmatrix} p & p/2 & p/3 & p/4 \\ p/2 & p & p/2 & p/3 \\ p/3 & p/2 & p & p/2 \\ p/4 & p/3 & p/2 & p \end{pmatrix} \tag{4}$$

Finally, rescale the probabilities in each row of the matrix for them to sum to 1.

- *Attenuation formula C.* Populate the matrix with p in the main diagonal. For the super-diagonal and the sub-diagonal raise p to the power of 1 plus the distance to the diagonal. Do the same for the super-super-diagonal and the sub-sub-diagonal, and so on. A 4×4 matrix would be as follows:

$$\mathbf{P} = \begin{pmatrix} p & p^2 & p^3 & p^4 \\ p^2 & p & p^2 & p^3 \\ p^3 & p^2 & p & p^2 \\ p^4 & p^3 & p^2 & p \end{pmatrix} \tag{5}$$

Finally, rescale the probabilities in each row of the matrix for them to sum to 1.

5.3 Analysis

Our analysis consists in estimating the empirical distribution $\hat{\pi}$ of the original data based on the randomized data by means of Expression (2), and then comparing $\hat{\pi}$ with the true distribution π of the original data. To transform continuous consumption data into categorical data amenable to randomization, we have split the range of consumption into 16 intervals. We have

performed 100 runs for each of the 9 RR probability matrices. Diagrams in Figures 3, 4 and 5 present results corresponding to the average of the 100 runs for each matrix. The abscissae in those diagrams represent the intervals, and the ordinates the relative frequencies of values that fall in each interval.

In each diagram the blue line shows the frequencies of π, the orange line the frequencies of $\hat{\pi}$ for $p \approx 0.4$, the gray line $\hat{\pi}$ for $p \approx 0.6$ and the yellow line $\hat{\pi}$ for $p \approx 0.8$. The values of p are approximate due to rescaling. In general, lower values of p offer more privacy, at the cost of less accuracy. It can be seen that $p = 0.6$ offers relatively high privacy, while maintaining accuracy at the same time. In fact, for the case of attenuation formulae A and B the difference the three values of p is small. However, for attenuation C, accuracy clearly increases with p.

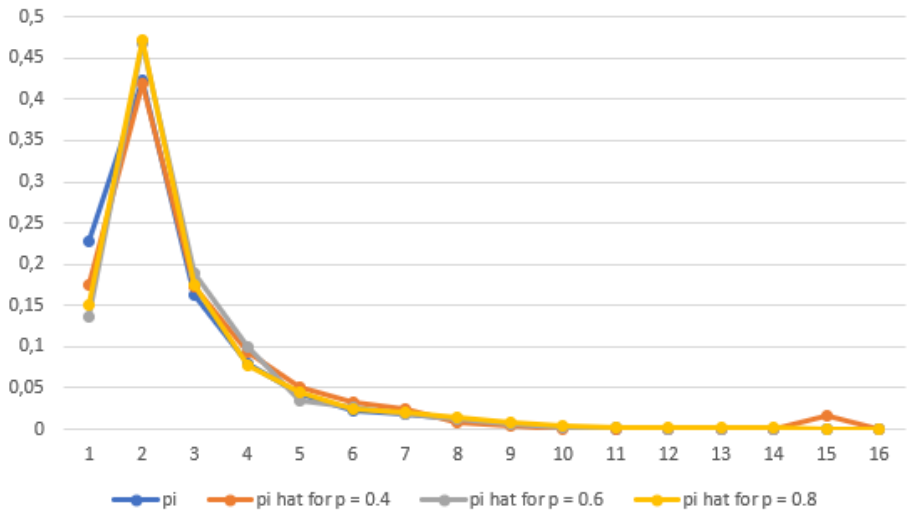

Figure 3: Attenuation A. Original frequencies and $\hat{\pi}$ frequencies for $p = 0.4$, 0.6 and 0.8

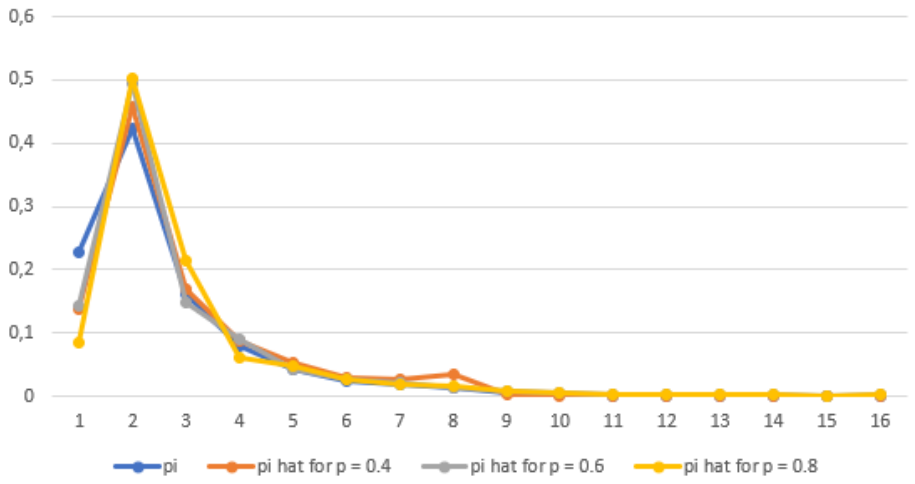

Figure 4: Attenuation B. Original frequencies and $\hat{\pi}$ frequencies for $p = 0.4$, 0.6 and 0.8

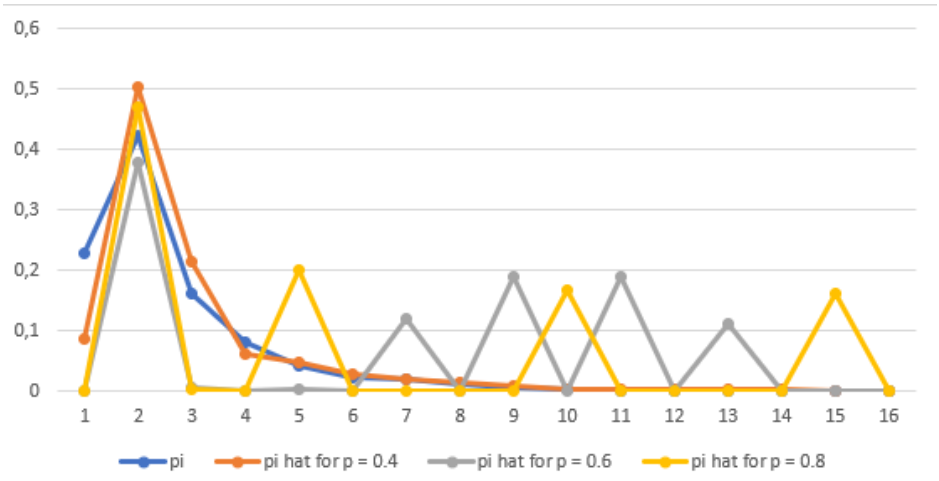

Figure 5: Attenuation C. Original frequencies and $\hat{\pi}$ frequencies for $p = 0.4$, 0.6 and 0.8

6 Conclusion and Future Research

Finding the appropriate probabilities for the RR matrix is a trade-off between privacy and accuracy. If the probability matrix is the identity matrix, where all values in the main diagonal equal to 1, we have 100% accuracy but no privacy. On the other hand, the more the probability matrix departs from the identity, the more privacy and the less accuracy.

As future work, we will quantify the privacy achieved in terms of differential privacy. Also, given a range for the SM readings, we will explore the best number of intervals to split that range, in order to achieve a reasonable trade-off between accuracy and computational complexity.

Acknowledgments and disclaimer

Thanks go to David Sánchez for comments on an earlier version of this manuscript. Partial support to this work has been received from the European Commission (project H2020-700540 "CANVAS"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705), and from the Spanish Government (project RTI2018-095094-B-C21). The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and do not necessarily reflect those of UNESCO.

References

[1] G. Ács and C. Castelluccia. I have a dream! (differentially private smart metering). In T. Filler, T. Pevný, S. Craver, and A. Ker, editors, *Information Hiding*, pages 118–132, 2011. Springer.

[2] R. Anderson and S. Fuloria. On the security economics of electricity metering. In *Workshop on Economics of Information Security - WEIS 2010*, 2010.

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

[4] H. Cao, S. Liu, Z. Guan, L. Wu, H. Deng, and X. Du. An efficient privacy-preserving algorithm based on randomized

response in IOT-based smart grid. In *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 881–886, Oct 2018.

[5] A. Chaudhuri and R. Mukerjee. *Randomized Response: Theory and Techniques*. Marcel Dekker, 1988.

[6] J. Domingo-Ferrer, R. Mulero-Vellido, and J. Soria-Comas. Multiparty computation with statistical input confidentiality via randomized response. In J. Domingo-Ferrer and F. Montes, editors, *Privacy in Statistical Databases*, pages 175–186, 2018. Springer International Publishing.

[7] J. Domingo-Ferrer and K. Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337-338:11 – 24, 2016.

[8] J. Domingo-Ferrer and J. Soria-Comas. Connecting randomized response, post-randomization, differential privacy and t-closeness via deniability and permutation. *CoRR*, abs/1803.02139, 2018.

[9] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, pages 1–12, 2006. Springer.

[10] Z. Erkin and G. Tsudik. Private computation of spatial and temporal power consumption with smart meters. In F. Bao, P. Samarati, and J. Zhou, editors, *Applied Cryptography and Network Security*, pages 561–577, 2012. Springer.

[11] F. D. Garcia and B. Jacobs. Privacy-friendly energy-metering via homomorphic encryption. In J. Cuellar, J. Lopez, G. Barthe, and A. Pretschner, editors, *Security and Trust Management*, pages 226–238, 2011. Springer.

[12] G. W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[13] T. Ryberg. *Smart Metering in Europe*, Berg Insight's M2M Research Series, Nov. 2018. `http://www.berginsight.com/ReportPDF/ProductSheet/bi-sm14-ps.pdf`, accessed 2019-04-22

[14] S. Sultan. Privacy-preserving metering in smart grid for billing, operational metering, and incentive-based schemes: A survey. *Computers & Security*, 84:148 – 165, 2019.

[15] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

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

[17] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.