Distributional Differential Privacy for Large-Scale Smart Metering

Márk Jelasity
University of Szeged, Hungary and MTA-SZTE
Research Group on Artificial Intelligence

Kenneth P. Birman
Cornell University, Ithaca, NY, USA

ABSTRACT

In smart power grids it is possible to match supply and demand by applying control mechanisms that are based on fine-grained load prediction. A crucial component of every control mechanism is monitoring, that is, executing queries over the network of smart meters. However, smart meters can learn so much about our lives that if we are to use such methods, it becomes imperative to protect privacy. Recent proposals recommend restricting the provider to differentially private queries, however the practicality of such approaches has not been settled. Here, we tackle an important problem with such approaches: even if queries at different points in time over statistically independent data are implemented in a differentially private way, the parameters of the distribution of the query might still reveal sensitive personal information. Protecting these parameters is hard if we allow for *continuous monitoring*, a natural requirement in the smart grid. We propose novel differentially private mechanisms that solve this problem for sum queries. We evaluate our methods and assumptions using a theoretical analysis as well as publicly available measurement data and show that the extra noise needed to protect distribution parameters is small.

1. INTRODUCTION

By deploying smart meters within individual homes and offices, it becomes possible to continuously measure, predict, and even control the consumption of power by the household. This could save money for consumers and for power producers, while also reducing unnecessary generation. An important component of any complete control solution that employs a network of smart meters is to monitor aggregate (predicted) consumption. Monitoring creates a challenge, however: we also need to ensure the privacy of the data, which can reveal the individual habits of the inhabitants of a home, reveal times when there is no one at home, the location of individuals within the home, or in extreme cases even very fine grained information such as which show is being watched on TV [1]. Our premise in this paper is that in an ideal system, personal data should be protected not only from eavesdroppers, but even from the utility itself.

The privacy protection problem has been well studied. The essential requirement is to calculate aggregation queries without revealing any individual records. Achieving this is non-trivial if the measurements are distributed: the meters should not trust one-another with sensitive data, nor can the network be trusted [2]. In addition, we must also make sure that the computed query results do not leak too much information about individual measurements either. The intent of the differential privacy model is to address this problem [3]. Differentially private query results contain a carefully designed amount of noise that masks the influence of any individual data record.

Unfortunately, supporting an unlimited number of queries introduces a further complication. Even if we implement a distributed differentially private mechanism, the stream of the query results still has a potential to leak information about the constant parameters of the distribution of the measurements in an individual home. That is, *existing techniques do not prevent information leakage about static properties* of the household, despite the fact that those static properties indirectly influence individual readings. This is a problem, because static properties that might be teased out over a period of time could still reveal sensitive information. Examples include the number of inhabitants, behavioral patterns and habits, the list of devices in the home, and so on.

In this paper we focus on solving the problem of protecting the privacy of distribution parameters even when an unlimited number of queries are allowed. Our main contribution takes the form of novel distributed differentially private mechanisms for sum queries that protect not only the individual records but also the parameters of individual energy consumption patterns. We evaluate our methods and assumptions using a theoretical analysis and publicly available measurement data. We believe ours is the first solution in which it is possible to monitor the network for an unlimited time while still achieving provable guarantees of differential privacy that extend to static aspects of personal information.

2. BACKGROUND AND RELATED WORK

It is possible to control energy consumption using smart devices with approaches ranging from purely local scheduling methods [4], to central control schemes that do nothing to protect privacy, to global self-organization without central control [5]. We are interested in the latter style of solutions. These typically start with a global aggregation component, and then use the output from the aggregation step as input to a control loop or a decision mechanism. As our global aggregation function, we focus on the sum function, which turns out to be a surprisingly powerful primitive both in and of itself, and for calculating more complex statistics [6, 7].

Preserving privacy in this context has been studied extensively. Cryptographic techniques have been proposed to perform computations in individual homes [8], for example, for the purposes of policy based billing. Our area of interest, aggregate computing, has also been addressed. One part of the problem is to be able to collaboratively compute the

IH&MMSec'14, June 11–13, 2014, Salzburg, Austria.

ACM 978-1-4503-2647-6/14/06 ...$15.00. http://dx.doi.org/10.1145/2600918.2600919.

sum of a set of values distributed over a network without any node revealing its value to any other node. Techniques for achieving this are known [2], and have typically been based on secret sharing schemes [9, 10].

The other part of the problem is to make sure that the computed aggregate query cannot be used to infer much information about individual records. *Differential privacy* (see also Section 5) is a framework for protecting data, whereby noise is judiciously introduced to the query result to mask the contribution and hence content of individual records [3]. Distributed implementations have been proposed that—on top of some secret sharing scheme for secure multi-party computations—also implement noise generation in a distributed way [11, 12]. We build on this work in that we will assume that an implementation of a distributed differentially private mechanism is available for computing sum queries and for generating Gaussian and Laplacian noise, and we will use these as black box components to implement our distributional differentially private schemes.

Our main focus here is on the problems that arise from the repeated computation of the sum query (as opposed to the single-shot approach of previous techniques). In fact there has been prior work on very closely related questions [13,14]. However, previously proposed approaches make very different kinds of assumptions than we do both about how the data is generated and exactly what needs to be protected. We revisit previous work in this area in Section 5.

3. PROBLEM STATEMENT

Let us assume that we have n smart meters. At a given time t, let the database containing the readings of the smart meters be $D(t) = (x_1(t), \ldots, x_n(t))$. These readings can correspond to actual power consumption at time t, aggregated power consumption in a short time interval before t, or predicted consumption in a short time interval after t. Since all these cases result in similar distributions, our discussion covers all these cases. Let us consider the series of databases $D_1, D_2, \ldots$, where $D_j = D(t_0 + j \cdot \delta)$. That is, the databases are defined as snapshots taken at regular time intervals starting at time t_0. Let us introduce the notation $D_j = (x_{1j}, \ldots, x_{nj})$.

The database D_j is distributed in that the smart meters do not upload their output to a central location. For the present paper, details of the distributed communication are irrelevant to the analysis: instead, we assume that there is a secure privacy preserving mechanism in place to compute the sum query of the readings, such as those discussed in [11, 12]. These mechanisms can deal not only with computing the sum query, but also with adding the necessary noise to it to achieve differential privacy in a fully distributed way. With this in mind, for the remainder of the treatment we will build on the primitives of summation and the addition of certain noise terms, noting that the actual implementation is intended to be fully distributed.

We assume that the databases are generated by some probability distribution, that is, we model each measurement x_{ij} as a random variable. We elaborate on this model in Section 4. The utility wishes to carry out a series of queries $M_j(D_j)$, with $j = 1, 2, \ldots$. Although the answers should become available to the utility in a timely manner, the computation must not reveal any of the individual values x_{ij}. In addition, the utility should not learn about the parameters of the underlying probabilistic models. As mentioned before, we focus on the sum query $M_j = \sum_{i=1}^{n} x_{ij}$.

4. A GENERATIVE PROBABILISTIC MODEL

Our model is shown in Figure 1(a). Variable M_j is the query result that is obtained over database D_j. In this model, we assume that the distribution of the variables x_{ij} depends on a set of global external parameters ϕ_j for all databases $j = 1, 2, \ldots$, and a set of internal smart meter parameters θ_i for all smart meters $1 \leq i \leq n$.

Parameters ϕ_j are common to all meters but depend on the time when the snapshot was taken. These include weather conditions, the day of week, public holidays, and the time of day as well. We assume that parameters ϕ_j are publicly known, and also that within a database D_i the values x_{ij} depend on each other only through the common parameters ϕ_j. This means that if the common parameters are known (as we have just assumed) then within any database D_j all variables x_{ij} are independent. Note that here we introduced a simplification by not considering any further structure, such as geography, that could result in parameters that are shared by a subset of meters. Every parameter is either fully local to a meter, or common to all meters.

Parameters θ_i are internal to smart meter i. As expressed by the plate model, these parameters are static during the observation of the meters. That is, these parameters are constant, but unknown. They describe, for example, the set of appliances in the home, and the stable habits, behavioral patterns, and preferences (that is, the personal profile) of the inhabitants. We want to make sure that the static parameters θ_i are also protected by a differentially private mechanism, in addition to the individual readings in the databases.

Note that—to simplify our discussion—we made the assumption that at any point in time the system follows some single underlying probabilistic model, but that the variables at different points in time are independent if the static parameters of the model are known. As it will be evident later, our approach to protect the static parameters θ_i is completely insensitive to this assumption but, in the lack of extra measures, the privacy of the individual readings x_{ij} could weaken if consecutive readings are correlated.

To characterize autocorrelation, we examined the publicly available SMART∗ dataset [15]. The dataset contains power consumption measurements in five second intervals in three homes that are called home A, B and C. The interval covered by the dataset spans from the 15th of April until the 5th of July, 2012. Due to the relatively limited amount of data that covers a relatively short amount of time, we considered only the time of day as a global parameter. Based on the available data, we tested the assumption of independence by extracting several time series that correspond to the same value of the global parameter, namely the time of day. More precisely, we divided the time into 30 minute intervals and aggregated the consumption in them. We then created a series using the values corresponding to the same time of day in the series of days that are covered in the dataset. Let such a series be denoted by x_{ij}^t where $i \in \{A, B, C\}$ selects the home, j selects the day, and t defines the time of day.

Figure 2 shows autocorrelation plots for the series $(x_{Aj})_{j=1}^{88}$, with $t \in \{\text{midnight}, 7\text{am}, 6\text{pm}\}$ (for homes B and C the plots are similar). The choice of t is arbitrary, but other values result in similar results. The data covers 88 consecutive days. We used 50 samples to calculate the approximation for each time lag to make sure the variance of the approximations is the same. Patterns in these graphs would indicate significant autocorrelation. None are evident, but notice that the confidence band is far from perfect, in that the underlying distribution is not normal (see Section 6.2).

While proposing a full control strategy is outside the scope of this paper, note that the above observation does not mean that control mechanisms that operate with a control period shorter than one day are impossible. Recall that we target communities with large numbers of consumers. In such a setting, we could introduce an increased amount of noise carefully calculated based on the observed correlations so as to protect x_{ij}. As we will see, the noise we need to add to each query has a small expected value independent of network

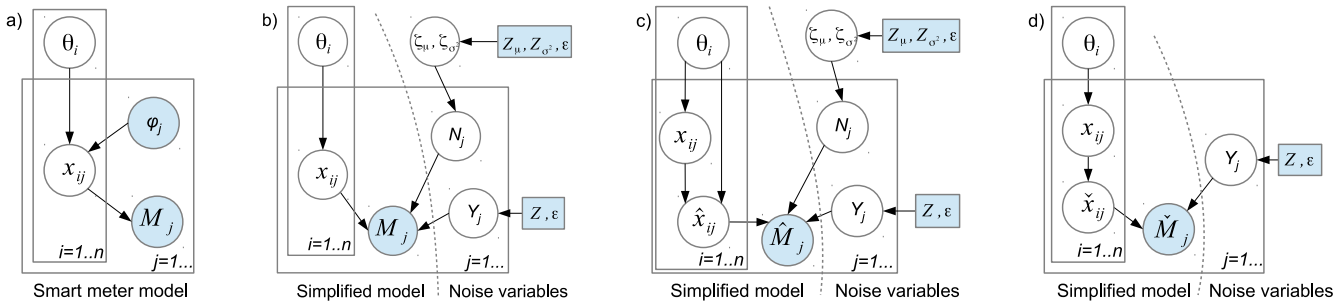

Figure 1: Probabilistic models of smart meter data and privacy mechanisms using plate notation. The shaded variables are known. The rest of the variables need privacy protection. a: our model of smart meters; b: model when x_{ij} is normally distributed; c: x_{ij} is not normally distributed and $\hat{x}_{ij}$ is normally distributed; d: arbitrarily distributed x_{ij} and $\check{x}_{ij} \sim$ Bernoulli(x_{ij}).

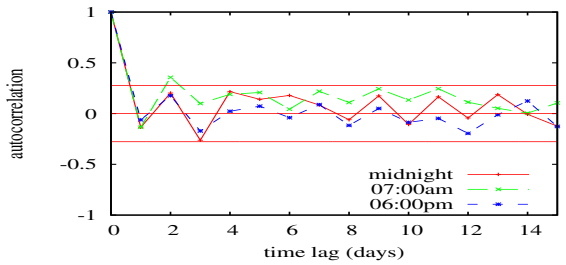

Figure 2: Autocorrelation plot of the time series of 30 minute measurement intervals at the same time of day during the days covered by the data. The 95% confidence interval assuming random data with normal distribution is shown.

size; but in a large network an increased (but still constant) amount of noise is still suitable. Also, to stabilize control, the provider might choose to sample the large network with each query. This also reduces correlation as a side-effect through the increased period of reading a particular meter. Moreover, sampling itself fits into our model, as taking a random subset simply introduces another private parameter: the probability whether the given meter is included or not.

5. DISTRIBUTIONAL PRIVACY

Let us now introduce the notion of differential privacy for general queries [16]. Let M be an algorithm producing an answer to a query issued on any possible database $D \in \mathcal{D}$. While operating on a fixed database D, algorithm M will also introduce random noise, thereby randomizing its output. That is, for a fixed database D, $M(D)$ will be a random variable. Let the distance function $d : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{N}$ be defined as the number of records in which two given databases differ. Without loss of generality, we assume that all the databases contain the same number of records.

Definition 1. (ε-*differential privacy*) Let M be a randomized mechanism acting on databases. M is ε-*differentially private* iff for any two fixed databases D and D' such that $d(D, D') = 1$, and for any output M, we have

$$P(D|M) \leq P(D'|M) \cdot \exp(\varepsilon). \tag{1}$$

We expressed the traditional definition in a Bayesian style. This definition is equivalent to the usual definition [16] if the prior distribution over the databases is uniform (any possible database is equally likely, that is, $P(D) = P(D')$).

One possible way to achieve differential privacy of the sum query in a database D_i is by adding noise to the output of the query calibrated according to the *sensitivity* of the query [17].

In other words, we can return

$$M_j = M(D_j) = Y_j + \sum_{i=1}^{n} x_{ij}, \tag{2}$$

where Y_j is an appropriate random variable. A common choice for the distribution of Y_j is $Y_j \sim$ Laplace$(0, Z/\varepsilon)$, where Z is a constant representing the *global sensitivity* of the sum function [3, 17]:

Definition 2. (*global sensitivity [17]*) The *global sensitivity* Z_f of $f : \mathcal{D} \mapsto \mathbb{R}$ is given by

$$Z_f = \max_{D,D':\, d(D,D')=1} |f(D) - f(D')| \tag{3}$$

We can apply this approach if there is a global upper bound on the values x_{ij}, since the global sensitivity of the sum is bounded by the maximal value of any addend. Such a bound can be assumed to exist in our application domain.

Now, we can turn to our goal, that is, to release M_j, $j = 1, 2, \ldots$. One possible approach would be to follow the work of Dwork et al [13], where the notion of event level privacy in *growing* databases is defined. Roughly speaking, the idea there is that two series are adjacent if they differ in one element. This is almost the same notion as the adjacency of databases, except that the static databases are now replaced by data streams. With this in mind, we could define a notion of adjacency of two series of databases by requiring that exactly one pair of databases is adjacent, and the rest of the pairs are identical (where identical time points define the pairs). We could then release M_j, $j = 1, 2, \ldots$, if the series is differentially private in terms of this adjacency definition. Indeed, the series M_j, $j = 1, 2, \ldots$ is ε-differentially private if all queries M_j are because of the *parallel composition* [18] of the queries (each query is run on a separate subset of the union of the available data).

As noted in the introduction, however, this form of protection is inadequate because it overlooks a potentially important form of leakage. Intuitively, even if all individual queries M_j are protected by adding noise, the parameters θ_i are still not protected if we can observe an unlimited number of query results. To see this, consider that if the global parameters ϕ_j are in fact constant (do not depend on j), then all variables M_j will have an identical distribution that can be recovered with an arbitrary precision after performing a sufficient number of measurements. This is because (as evident from Figure 1(a)) in this case all variables x_{ij} will have the same distribution for all j. In the case of the sum query, and if the sensitivity mechanism is applied, this distribution can be determined from equation (2). Since Z and ε are known, as are the ϕ_j, the unknown parameters are θ_i. This means that approximating the query distribution could lead to information leakage about the private parameters θ_i of the smart meters.

If the global external parameters ϕ_j are time-varying, then the parameters θ_i would be more secure, since in that case the shape of the distribution of the queries will be more complex (it will become a mixture distribution determined by the distribution of ϕ_j). An adversary interested in θ_i will therefore attempt to limit itself to considering only subsets of the readings that share the same external parameters, since that way approximating θ_i is easier. In other words, a constant ϕ_j is the worst case for privacy. The considerations above motivate the following definition of adjacency.

Definition 3. (*distributional adjacency*) Let us assume the probabilistic model in Figure 1(a). Consider two series of databases $(D_j)_{j=1}^{\infty}$ and $\left(D'_j\right)_{j=1}^{\infty}$ that were generated by the model. The two series are *distributionally adjacent* iff $\theta_i = \theta'_i$ for all but one index $1 \leq k \leq n$, for which $\theta_k \neq \theta'_k$, and all the other variables that do not depend on θ_k are the same.

The intuition behind the definition is simple. When monitoring smart meters, distributional adjacency captures the situation when we are collecting smart metering data in a set of homes that differs in exactly one element (we replace one home with another one), but otherwise everything remains exactly the same including all the readings in the rest of the homes. Based on this notion of adjacency, we define distributional differential privacy.

Definition 4. (*distributional ε-differential privacy*) Let M be a randomized mechanism acting on databases. Let us assume the probabilistic model in Figure 1(a). M is *distributionally ε-differentially private* iff for any two fixed and distributionally adjacent database series $(D_j)_{j=1}^{\infty}$ with parameters $\theta = (\theta_1, \ldots, \theta_n)$ and $\left(D'_j\right)_{j=1}^{\infty}$ with parameters $\theta' = (\theta'_1, \ldots, \theta'_n)$, and for any query output series $(M_j)_{j=1}^{\infty}$

$$P(\theta|\,(M_j)_{j=1}^{\infty}) \leq P(\theta'|\,(M_j)_{j=1}^{\infty}) \cdot \exp(\varepsilon). \tag{4}$$

6. ACHIEVING DISTRIBUTIONAL PRIVACY

Here, we provide techniques to achieve distributional privacy that differ in their assumptions about the available knowledge and the distribution of x_{ij}. First, let us ignore the parameter ϕ_j in the model in Figure 1(a). As we argued before, this is the worst case from the point of view of privacy, since the adversary has a noise-free sample of the distribution of the query. Besides, in a real system an adversary can collect samples that belong to the same value of ϕ_j; recall that the value of ϕ_j is known publicly.

6.1 Readings with Gaussian distributions

We provide an example where distributional ε-differential privacy can be achieved. Let us assume that measurement x_{ij} has a Gaussian distribution: $x_{ij} \sim \mathcal{N}(\theta_i)$ for all j, where $\theta_i = (\mu_i, \sigma_i^2)$. In this case, assuming the model in Figure 1(a) (without ϕ_j), we know that $\sum_{i=1}^n x_{ij} \sim \mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$. In other words, the sum query that we are interested in has a Gaussian distribution as well. Many other distributions have a similar property, for example, the binomial or the Poisson distributions. The class of *stable* distributions also has this property. This class covers many practical distributions including normal and power-law distributions [19]. Most importantly, in this case the distribution of the query is simple and it has only a few parameters that depend on the parameters of the distributions of the individual measurements.

Observe that by returning an infinite series of query results the mechanism essentially returns the sum query over the parameters of the distributions θ_i. We never obtain the exact parameter set of the query distribution (the sum of θ_i), but instead we can draw an unlimited number of samples from this distribution. Even so, let us make the very conservative assumption that eventually the adversary learns the exact parameters in this case; doing so only makes it harder to achieve distributional privacy.

To make the solution distributionally private, we can apply, among other options, a sensitivity-based approach to the parameter space. That is, we can add independent noise $N_j \sim \mathcal{N}(\zeta_\mu, \zeta_{\sigma^2})$, which results in the query distribution

$$N_j + \sum_{i=1}^n x_{ij} \sim \mathcal{N}(\zeta_\mu + \sum_{i=1}^n \mu_i, \zeta_{\sigma^2} + \sum_{i=1}^n \sigma_i^2), \tag{5}$$

where ζ_μ and ζ_{σ^2} are constants that are drawn from the distribution Laplace$(0, Z_\mu/\varepsilon)$ and Laplace$(0, Z_{\sigma^2}/\varepsilon)$, respectively, where Z_μ and Z_{σ^2} are the global sensitivities of the sum queries $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$, respectively. Constants ζ_μ and ζ_{σ^2} are drawn at the beginning of collecting the query results and are not changed later. This, similarly to traditional differential privacy, results in a noisy result; but this time this noise will be applied to the parameters of the distributions, and not the data.

Indeed—under the assumption that the adversary will eventually learn from the infinite series of query results the exact parameters $(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$—distributional privacy (Definition 4) becomes simply an instance of differential privacy (Definition 1) with the database being the distribution parameters (μ_i, σ_i^2), $i = 1, \ldots, n$ over which a differentially private sum query is run. This proves that the mechanism in equation (5) is distributionally ε-differentially private.

Finally, to also achieve the differential privacy of the data in each individual database in time, we introduce the usual noise term, as mentioned previously in equation (2):

$$M_j = N_j + Y_j + \sum_{i=1}^n x_{ij}, \tag{6}$$

where $Y_j \sim$ Laplace$(0, Z/\varepsilon)$ and Z is the global sensitivity of the sum query. This will not change distributional differential privacy, since adding additional independent noise will never weaken the privacy of any scheme. Also, note that $Z \geq Z_\mu$. The resulting model is illustrated in Figure 1(b).

We stress that in this example we assumed that an unlimited number of samples are available from the same query distribution, and so the parameters of the query distribution can be recovered to an arbitrary precision. Nonetheless, we were able to achieve distributional differential privacy due to the special property of Gaussian distributions.

6.2 Realistic distributions

The distribution of the readings is not normal (and not even stable) in practice. Indeed, Figure 3 (left) illustrates the probability density of power consumption as a function of the time of day in home A in the SMART∗ dataset [15]. For this plot we aggregated the consumption in 5 minute intervals and produced a scatterplot based on the days that are covered in the dataset. These plots (and related work [20]) suggest that the distribution of power consumption at a certain time of day is a mixture distribution. In a mixture distribution the variables that select the components of the mixture are those internal variables that are not constant during the observation period (for example, whether the owner is on a holiday or not, whether there are guests in the home, whether the air conditioning is on, etc.).

In the case of mixture distributions, the number of parameters of the distribution of the query will grow with the number of readings used to answer the query, unlike in the case of stable distributions mentioned in Section 6.1. This creates a new challenge to distributional privacy, particularly given our assumption that the adversary can recover the exact parameters of the query distribution using the unlimited

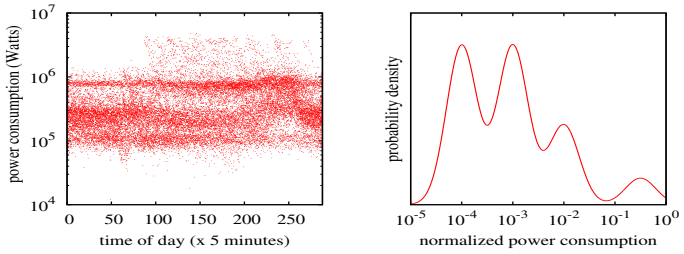

Figure 3: Left: scatter plot to illustrate the observed probability density of energy consumption as a function of time of day, based on 5 minute intervals during the days covered by the data. Right: probability density for modeling power consumption prediction. This is a lognormal mixture distribution over the $[0, 1]$ interval.

number of query samples. In practice an additional source of uncertainty regarding the parameters of the distribution is the limited number of samples available; a topic we do not exploit in this paper.

6.3 Transformation to Gaussian distributions

Our first approach to tackling arbitrary distributions converts the meter readings' distributions to Gaussian. The main advantage of this approach is that we can theoretically guarantee distributional privacy for arbitrary distributions for the case of the sum query, while introducing negligible extra noise. However, we need to assume that the local parameters θ_i are known locally by meter i. This assumption is reasonable, since the local meter has full access to local consumption data and hence can easily approximate the distribution. Further, if we aggregate predicted consumption, then the parameters are fully determined by the local meter.

The idea is that, based on the knowledge of θ_i and x_{ij}, the local meter i will generate a variable $\hat{x}_{ij}$ that has a Gaussian distribution with the same expectation and standard deviation as x_{ij} and is maximally correlated with x_{ij}. After this, the query is calculated using $\hat{x}_{ij}$ instead of x_{ij} using the same noise variables as in equation (6):

$$\hat{M}_j = N_j + Y_j + \sum_{i=1}^{n} \hat{x}_{ij}, \tag{7}$$

while keeping x_{ij} private. Applying the same reasoning as in Section 6.1, it is clear that the method is distributionally differentially private. Instead of a Gaussian distribution, any other stable distribution can be used, as mentioned in Section 6.1, depending on the shape of the distribution to be approximated.

The resulting probabilistic model is illustrated in Figure 1(c). Let us elaborate on how $\hat{x}_{ij}$ is computed. Let (μ_i, σ_i^2) be the expectation and variance of x_{ij}, and let $\mathcal{X}(x) = P(x_{ij} \leq x)$ be the distribution function of x_{ij}. We know that $\mathcal{X}(\cdot)$ and (μ_i, σ_i^2) are known locally. Let

$$\hat{x}_{ij} = \mathcal{N}^{-1}(\mathcal{X}(x_{ij}); \mu_i, \sigma_i^2), \tag{8}$$

in other words, we compute $\hat{x}_{ij}$ in such a way that it corresponds to the same quantile according to $\mathcal{N}(\cdot; \mu_i, \sigma_i^2)$ as x_{ij} according to $\mathcal{X}(\cdot)$. (To simplify the discussion, without loss of generality, we assumed that $\mathcal{X}(\cdot)$ is continuous.)

It is obvious that the expectation and variance of M_j and $\hat{M}_j$ are the same since x_{ij} and $\hat{x}_{ij}$ have the same expectation and variance by design for all (i, j), the readings x_{ij} are independent (and hence the transformed readings $\hat{x}_{ij}$ are independent as well), and M_j and $\hat{M}_j$ are defined by the same linear function of the readings. Furthermore, the distributions of M_j and $\hat{M}_j$ will be very similar. This is because $\hat{M}_j$ is normally distributed, while M_j is the sum of many similar variables, so it is also likely to be normally distributed.

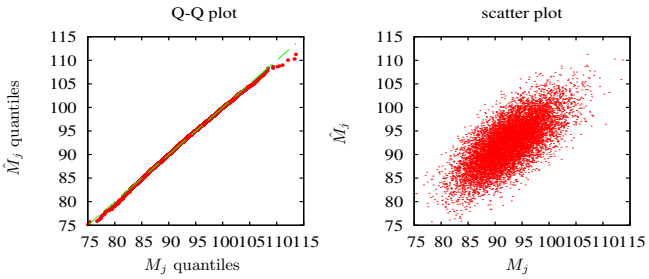

Figure 4: Q-Q plot (left) and scatter plot (right) based on 10,000 observations. The correlation coefficient is 0.7.

To examine the distribution of $(M_j, \hat{M}_j)$ empirically, we model the distribution of x_{ij} at time j for all i using the mixture distribution in Figure 3 (right). This distribution is an approximation of the distribution in home A at noon (Figure 3 (left)). We assume that there are 1000 meters (i.e., $i = 1, \ldots, 1000$). We empirically generate 10,000 independent samples of $(M_j, \hat{M}_j)$. The quantile-quantile (Q-Q) plot in Figure 4 clearly shows that M_j is almost exactly normal. In addition, the scatter plot in Figure 4 shows a high correlation. Overall, we can conclude that in a realistic setting the proposed transformation preserves most of the original information while providing full distributional privacy.

6.4 Transformation to Bernoulli distributions

Our second approach to tackle arbitrary distributions is based on converting the meter readings' distributions to a Bernoulli distribution. The main advantage of this approach is that we will *not* assume that the local parameters θ_i are known locally by meter i. Also, we will theoretically prove that this approach will protect the privacy of θ_i, with the exception of the expected value $E(x_{ij})$. The approach is also very simple to implement. However, this approach introduces a higher level of noise, as we will see.

Let us assume that the distribution of the reading x_{ij} is arbitrary, but the values are bounded. Without loss of generality, let $0 \leq x_{ij} \leq 1$. Let us introduce a new variable $\check{x}_{ij}$ for all readings x_{ij} where $\check{x}_{ij} \sim \text{Bernoulli}(x_{ij})$.

We calculate the query over the new variables $\check{x}_{ij}$ while the variables x_{ij} are kept private. Since $E(\check{x}_{ij}|x_{ij}) = x_{ij}$, we know that $E(\check{x}_{ij}) = E(E(\check{x}_{ij}|x_{ij})) = E(x_{ij})$, which means that, due to the linearity of expectation and the construction of $\check{x}_{ij}$, $E\left[\sum_{i=1}^{n} \check{x}_{ij}\right] = E\left[\sum_{i=1}^{n} x_{ij}\right]$. That is, using variables $\check{x}_{ij}$ results in the same expected query value for the sum query and, in fact, for any linear query. The same observations hold also if we consider index j and fix i. Further, and most importantly, the series $(\check{x}_{ij})_{j=1}^{\infty}$ carries no information about the parameters of the distribution of x_{ij} other than the expected value, since all the values are 0 or 1, and they are drawn independently.

As a practical technique, we propose to return the query

$$\check{M}_j = Y_j + \sum_{i=1}^{n} \check{x}_{ij}. \tag{9}$$

Here—as before—$Y_j \sim \text{Laplace}(0, Z/\varepsilon)$ and Z is the global sensitivity of $\sum_{i=1}^{n} \check{x}_{ij}$. Clearly, due to the Bernoulli distribution of $\check{x}_{ij}$ we have $Z = 1$. Since the authors are not aware of any closed form for the convolution of multiple Bernoulli distributions with different parameters, the Bernoulli variables are not made distributionally private here. However, since these variables only reveal the expectation of the com-

mon distribution of the masked variables x_{ij}, we protect most of the fine structure of the local distribution. The resulting probabilistic model is illustrated in Figure 1(d).

Let us examine exactly how much noise we introduced. Our first intuition is that in a usual setting this noise is in the same order of magnitude as the noise introduced by sampling x_{ij} using parameters θ_i. Additionally, this noise will also decrease in a relative sense as the number of smart meters increases. More precisely, the variance of the distribution Bernoulli(p) is $p(1-p)$. This is maximal if $p = 0.5$. Now, under the probabilistic model we work with this means that

$$\text{stdev}\left[\sum_{i=1}^{n} \tilde{x}_{ij}\right] = \sqrt{\sum_{i=1}^{n} x_{ij}(1-x_{ij})} \leq \frac{\sqrt{n}}{2} \qquad (10)$$

since the variables are independent. The worst case of $\sqrt{n}/2$ is given when $x_{ij} = 0.5$ for all i.

This noise, however, is not necessarily extra noise from the point of the view of the application. For example, if the variables x_{ij} are statistical predictions then the question is the ratio of the expected noise that originates from the uncertainty in the prediction and the extra noise due to converting the variables. To examine this case, as in Section 6.3, we again model the distribution of the prediction using the mixture distribution in Figure 3. As before, the consumption values are normalized to the interval $[0, 1]$. Setting $n = 1000$ and taking 10,000 samples of $\sum_{i=1}^{1000} \tilde{x}_{ij}$ we find the empirical standard deviation to be 9.15. At the same time, $\text{stdev}(\sum_{i=1}^{1000} x_{ij}) = 4.82$, which gives the ratio of 1.9. This ratio is constant as a function of n, and both standard deviations are $O(\sqrt{n})$. For the sake of completeness, for $n = 1000$, the upper bound of standard deviation is $\sqrt{1000}/2 = 15.81$. We achieve a variance of 9.15 due to the strong asymmetry of the original distribution.

7. CONCLUSIONS

We proposed novel techniques to implement distributed sum queries in a privacy preserving way. We believe that our work is the first to offer practical options for achieving full privacy covering both individual readings and hidden static parameters. The key insight was that if the individual meter readings (or predictions) are normally distributed then the sum query will also be normally distributed. This allows us to apply differentially private techniques on the distribution parameters. Since normality is not always satisfied, we proposed techniques to transform the distributions. We argued that the extra noise due to these techniques is small. In a full practical implementation of our distributed sum queries the noise terms we identified and the sum itself can be computed in a distributed and private way, a problem known to be tractable [11, 12]. A full control solution based on the monitoring approach described here is the subject of our ongoing research.

8. ACKNOWLEDGMENTS

This work was supported, in part, by grants from the US NSF, from the ARPAe "GENI" program at DOE, and from the EU and the European Social Fund through project FuturICT.hu (grant no .: TAMOP-4.2.2.C-11/1/KONV-2012-0013). M. Jelasity was supported by the Fulbright Program and the Bolyai Scholarship of the Hungarian Acad. Sci.

9. REFERENCES

[1] Rouf, I., Mustafa, H., Xu, M., Xu, W., Miller, R., Gruteser, M.: Neighborhood watch: security and privacy analysis of automatic meter reading systems. In: Proc. 2012 ACM Conf. on Comp. and Comm. Security (CCS'12), ACM (2012) 462–473

[2] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, M.Y.: Tools for privacy preserving distributed data mining. SIGKDD Explor. Newsl. **4**(2) (2002) 28–34

[3] Dwork, C.: A firm foundation for private data analysis. Commun. ACM **54**(1) (January 2011) 86–95

[4] Barker, S., Mishra, A., Irwin, D., Shenoy, P., Albrecht, J.: Smartcap: Flattening peak electricity demand in smart homes. In: IEEE Intl. Conf. on Pervasive Computing and Comm. (PerCom). (2012) 67–75

[5] Beal, J., Berliner, J., Hunter, K.: Fast precise distributed control for energy demand management. In: IEEE Sixth Intl. Conf. on Self-Adaptive and Self-Organizing Systems (SASO). (2012) 187–192

[6] Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the sulq framework. In: Proc. 24th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS'05), ACM (2005) 128–138

[7] Zhang, J., Zhang, Z., Xiao, X., Yang, Y., Winslett, M.: Functional mechanism: regression analysis under differential privacy. Proc. VLDB Endow. **5**(11) (July 2012) 1364–1375

[8] Rial, A., Danezis, G.: Privacy-preserving smart metering. In: Proc. 10th annual ACM workshop on Privacy in the electronic society (WPES'11), ACM (2011) 49–60

[9] Maurer, U.: Secure multi-party computation made simple. Discrete Applied Math. **154**(2) (2006) 370–381

[10] Yao, A.C.C.: How to generate and exchange secrets. In: Proc. 27th Annual Symposium on Foundations of Comp. Sci. (FOCS). (October 1986) 162–167

[11] Ács, G., Castelluccia, C.: I have a dream! (differentially private smart metering). In: Information Hiding. LNCS 6958. Springer (2011) 118–132

[12] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Advances in Cryptology - EUROCRYPT 2006. LNCS 4004. Springer (2006) 486–503

[13] Dwork, C., Naor, M., Pitassi, T., Rothblum, G.N.: Differential privacy under continual observation. In: Proc. 42nd ACM symposium on Theory of computing (STOC'10), ACM (2010) 715–724

[14] Ny, J.L., Pappas, G.J.: Differentially private filtering. Technical Report 1207.4305, arxiv.org (2012)

[15] Barker, S., Mishra, A., Irwin, D., Cecchet, E., Shenoy, P., Albrecht, J.: Smart*: An open data set and tools for enabling research in sustainable homes. In: Proc. 2012 Workshop on Data Mining Applications in Sustainability (SustKDD 2012). (August 2012)

[16] Dwork, C.: Differential privacy. In: Automata, Languages and Programming (ICALP). LNCS 4052. Springer (2006) 1–12

[17] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography. LNCS 3876. Springer (2006) 265–284

[18] McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proc. 2009 ACM SIGMOD Intl. Conf. on Mgmt. of Data (SIGMOD'09), ACM (2009) 19–30

[19] Nolan, J.P.: 1. In: Stable Distributions - Models for Heavy Tailed Data. Birkhauser (2013) to appear.

[20] Barker, S., Kalra, S., Irwin, D., Shenoy, P.: Empirical characterization and modeling of electrical loads in smart homes. In: The Fourth Intl. Green Computing Conf. (IGCC'13). (2013)