National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

01-19

Protecting the Privacy of Smart Meter Data: the Differential Privacy Approach and the Multiplicative Noise Approach

John Brakenbury, P.Y O'Shaughnessy, Yan-Xia Lin

Copyright © 2019 by the National Institute for Applied Statistics Research Australia, UOW.
Work in progress, no part of this paper may be reproduced without permission from the Institute.

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Protecting the Privacy of Smart Meter Data: the Differential Privacy Approach and the Multiplicative Noise Approach

John Brackenbury*, **P.Y. O'Shaughnessy***, **Yan-Xia Lin***

*School of Mathematics and Applied Statistics, University of Wollongong, NSW, 2522, Australia

E-mail: poshaugh@uow.edu.au, yanxia@uow.edu.au

Abstract. Smart meter electricity data presents a risk of malicious agents gaining insight of private information, including residents' lifestyle and daily habits. In this paper, we consider two approaches to the masking of smart meter data to minimise disclosure of a dwelling's consumption signal to third parties including an energy provider, whilst enabling the sum of a cluster of households to be estimated properly. The first approach, DREAM is an existing differential privacy method which is based on additive noise and encryption. We propose an alternative data masking method using the 'Twin Uniform' multiplicative noises. Analytical results are derived for each scheme, followed by comparison through real smart meter data from ESSnet Big Data. We conclude that the Multiplicative Noise method preforms comparably to the existing DREAM method, while it holds its advantage in simple implementation.

Keywords. Differential Privacy, Multiplicative Noise, Additive Noise Method, Masked Data, Disclosure Risk, Utility Loss

1 Introduction

Given the rapid growth in big data analysis as a tool for commercial insight, many organisations often outsource the analyses of the data they collect in-house to a third-party firm specialising in data analysis. Indeed, even when the analysis is performed in-house, the initial phase of software development may require third-party developers who potentially have access to sensitive data. In other cases, institutions may benefit from the external data of customers from other institutions, such as in the area of fraud detection in the finance sector, but this presents a potential for breaching of privacy.

Whilst the rise of big data has created an entirely new industry in the past few decades, the handling of sensitive data presents a privacy concern for individuals. Notwithstanding the legal nuances of privacy policies, there is an expectation among consumers that their sensitive data would be protected, including a level of anonymisation when transferring control of data to third parties. This is indeed a challenge for businesses desiring insights from data: how can a firm meet its need for accurate and meaningful analysis while protecting individual's need for privacy? A helpful illustration is provided by Laforet et al. [9].

One of the areas of application is the 'smart meter', which is an emerging technology in the energy industry that records time-series electricity consumption data in contrast to the conventional cumulation methodology. Smart meter uses an advanced signal analysis technique called non-intrusive load monitoring (NILM) [6], dating back at least to 1992. It involves the analysis of changes in voltage and current entering a dwelling to determine the use of power, with potential accuracy up to distinguishing individual appliances. The meter transmits consumption figures automatically at regular intervals, which allows for more accurate billing with minimal estimation required. In addition, it eliminates the requirement of company workmen to physically visit locations for measurements. Such time-series data also provides a more high resolution view of demand for individual dwelling, enabling optimisation of power supply around peak periods. However, the time-series data produced by smart meters also provides observers opportunity to obtain sensitive information regarding the dwelling, including the number of occupants and their daily schedule.

In addition to the need of the customers, the needs of the energy providers must also be considered. For accurate billing, the *total consumption* over the billing period calculated from the altered time-series must not deviate beyond a certain tolerance from the true total. When the company intends to analyse network demand throughout the day, such analysis will not be meaningful unless the deviation on *each time point* is limited.

The question of how to measure protection procedures and assess data privacy is broad, with a vast amount of literature on the topic. Given the diversity of data types and purposes, privacy metrics are often chosen according to application, making comparison across applications problematic. Thought must always be given to the nature of what is being protected, and to the means by which unauthorised agents may access such information.

In 2010, Bohli, Sorge and Ugus published a paper specifically addressing the privacy concerns of smart meters [2]. Whilst recognising the needs for individual billing, and network demand, the authors emphasised that even the electricity suppliers operating the meters should be kept from possessing sensitive data pertaining to individuals' lifestyles. The authors suggest that aggregation such as summing along the time-series could be performed by the smart meter itself, given the meters are considered trusted devices without the risk of privacy compromise to both the electricity suppliers and any third-party. This is certainly a feasible option given current hardware.

In 2011, Ács and Castelluccia [1] proposed a method of data privacy for smart meters, DiffeRentially privatE smArt Metering (DREAM). Based on the concept of differential privacy, a perturbative scheme is proposed that ensures the privacy of individual time-series transmitted to an electricity supplier for aggregate demand analysis at each time point. It is assumed that no individual time-series are needed, but that the sum of sufficiently small clusters of households at each time point is sufficient resolution. Each meter applies a Gamma noise to its own signal, resulting in Laplace noise when cumulated across the cluster. Thus, after decryption using shared keys, the electricity supplier is only able to determine a noisy sum from each cluster, with each individual household differentially protected.

The authors took a holistic approach to their research, considering not only the statistical aspects, but also the technical practicality of encryption, transmission and meter capabilities. They provide a detailed assessment of potential attacks, including the results of numerical attempts at analysing the masked data using different inference techniques. The scheme proposed unfortunately does not allow legitimate analysis of microdata from individual dwellings. It also suffers from significant vulnerability to meter failure due to the nature of its encryption.

One of the challenges for the smart meter is missing data caused by meter failure. In this article, we propose to use a method involving multiplicative noise to tackle the issues caused by the missing data as well as to retain the privacy requirement for smart meter data. Noise multiplication is not a new data masking method. However to the best of our knowledge, it has not been implemented in the smart meter data, which has its unique requirement for privacy. We introduce a real smart meter dataset from European Statistical System and discuss the relevant testing criteria for measuring privacy in Section 2. Section 3 describes the existing method for the differentially private smart metering, which uses differential privacy method to mask smart meter data, and we propose a new data masking technique using multiplicative noise in Section 4. Section 5 compares the performance of the two data masking techniques in three aspects using the smart meter dataset and the conclusion is given in Section 6.

2 Data Description and Testing Criteria

In this section, we describe a real dataset for illustrating different data masking methods for smart electricity meter in this paper, and introduce the testing criteria for measuring the performance of the various techniques. The smart meter data used in this paper was acquired from ESSnet Big Data[1], a current project within the European Statistical System (ESS) that has been operating from February 2016. The data is recorded between 2013 and 2015 in Denmark and Estonia. Denmark and Estonia were selected for the study simply for their lack of legal barriers to the data collection, as well as the pre-existence of smart meter data suitable for the project's purposes. After data cleaning, $N = 3371$ series of 4-day long hourly ($T = 96$ time points) electricity consumption data was produced with entries measured in kilowatt-hours (kWh).

Let X_t denote the smart meter readings time t for $t = 1, \ldots, T$. More specifically, we assume that for $i = 1, \ldots, N$, the smart meter readings for the ith household at time t, X_t^i, follows the same distribution as X_t. This data set is used to compared to determine the utility and protective efficacy of various data-masking methods in the specific application of protecting individual households' electricity data. In particular, we seek to simultaneously

- enable accurate estimation of the sum $S_t^c = \sum_{i=1}^{n_c} X_t^i$ for a defined cluster of n_c dwellings at each time t; and
- protect the individual values $\{X_t^i\}$ from accurate estimation by the electricity suppliers .

The N households are separated into C clusters and each cluster has n_c households. We define S_t to be aggregated sum for a cluster of smart meter readings at time t, which defines the underlying distribution for S_t^c for $c = 1, \ldots, C$.

We introduce three testing criteria based on the relative error. Note that comparison in terms of absolute error is not sensible, due to the high variability of magnitude of the signals across different households and different time points. Error of 0.1 KWh may be allowable for a household averaging 1.8 KWh at each time point but unacceptable for a household averaging 0.5 KWh at each time point. Therefore absolute scale measures such as standard deviation cannot be used. Instead, we use the relative error of the estimates, which is often simply referred to as 'error' for brevity throughout this paper.

[1]See `ebgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data` for more information.

The first measure is the *probability of relative error below a tolerance threshold*,

$$p_{\delta,S}(t) = P\left(\left|\frac{\hat{S}_t - S_t}{S_t}\right| < \delta\right), \qquad p_{\delta,X}(t) = P\left(\left|\frac{\hat{X}_t - X_t}{X_t}\right| < \delta\right),$$

where $\delta > 0$ is the tolerance threshold. We denote $\hat{S}_t$ and $\hat{X}_t$ as the estimated values for cluster sum at time t and individual household at time t, respectively. The calculations of $\hat{S}_t$ and $\hat{X}_t$ vary according to the implemented data masking methods, estimated from the masked values received. Referred to as the *rate of error* or *accuracy rate*, p_δ in both cases are estimated by the sample proportions

$$\hat{p}_{\delta,S}(t) = \frac{1}{C}\sum_{c=1}^{C} I\left(\left|\frac{\hat{s}_t^c - s_t^c}{s_t^c}\right| < \delta\right), \qquad \hat{p}_{\delta,X}(t) = \frac{1}{N}\sum_{i=1}^{N} I\left(\left|\frac{\hat{x}_t^i - x_t^i}{x_t^i}\right| < \delta\right),$$

where $I(\cdot)$ denotes the indicator function and the lower cases x and $\hat{x}$ are the actual observations of variables X and $\hat{X}$, respectively. The cluster sum $s_t^c = \sum_{i=1}^{n_c} x_t^i$ is the realizations of S_t^c. The estimated masked cluster sum $\hat{s}_t^c$ is similar.

We use $p_{\delta,S}(t)$ and $p_{\delta,X}(t)$ to assess the accuracy of $\hat{S}_t$ and $\hat{X}_t$, respectively, but our goal in each context is opposite to each other. For estimation of the sum S_t we wish high accuracy to maximise utility, corresponding to high $p_{\delta,S}(t)$. For the individual household values X_t we wish to maximise privacy and minimise malicious estimation accuracy, corresponding to low $p_{\delta,X}(t)$. The latter is not relevant for DREAM, in which the micro data is not available, but is critical in measuring the vulnerability of the newly proposed Multiplicative scheme .

The second and third measures used in this paper are *mean relative error* (MRE) and *mean unsigned relative error* (MURE), namely

$$\text{MRE} = \frac{1}{C}\sum_{c=1}^{C}\frac{\hat{s}_t^c - s_t^c}{s_t^c}, \qquad \text{MURE} = \frac{1}{C}\sum_{c=1}^{C}\left|\frac{\hat{s}_t^c - s_t^c}{s_t^c}\right|.$$

Distinct from the rate of error $p_{\delta,S}$, MRE indicates the overall or average relative error. Since this includes both overestimation and underestimation, the errors of opposite sign offset each other, obscuring the magnitude of the relative errors themselves but giving an idea of the overall bias at each time point. MURE eliminates sign and the accompanying offset, providing a better measure of the average magnitude of relative error.

Another measure of risk used in assessing the Multiplicative scheme is the *correlation* between the malicious estimates and their true values, $\text{corr}(\hat{Y}_t, Y_t)$, where $Y_t = X_t + a$ is the shifted true value. This is utilised as a metric of regression risk, the disclosure risk posed by a malicious agent improving their estimation by use of a regression model. More details are given in Section 4.

3 DREAM: DiffeRentially privatE smArt Metering

3.1 Scheme

DiffeRentially privatE smArt Metering (DREAM) assumes N smart meters across an electricity distribution network, recording consumption data across regular time intervals as a time-series [1]. The smart meter is assumed a trusted device in this scenario, meaning it

will only operate as intended and is impervious to malicious interference, such as hacking to send data elsewhere. The paper defines each smart meter as a node, thus we follow the terminology here. The ith meter's (node) measurement at time t is denoted X_t^i, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The following is a simplification of the procedure, omitting some details regarding modular arithmetic and key selection.

Algorithm: DREAM

1. At each time point, each node adds Gamma difference noise to the signal: $\tilde{X}_t^i = X_t^i + G_1(1/N, 1/\lambda) - G_2(1/N, 1/\lambda)$.
2. The nodes randomly pair up, and each pair (i, j) produces a private encryption key $k_{i,j}$. Neither the pairing nor the encryption key is known by the supplier.
3. The encrypted, noisy signal $Enc(\tilde{X}_t^i) = X_t^i + G_1(1/N, 1/\lambda) - G_2(1/N, 1/\lambda) + k_{i,j}$ is then sent to the supplier by the node.
4. The supplier sums the signals across a cluster of n_c household and obtains the aggregate cluster sum with Laplace noise for the cth cluster,

$$\begin{aligned}\hat{S}_t^c &= \sum_{i=1}^{n_c} Enc(\tilde{X}_t^i) = \sum_{i=1}^{n_c} X_t^i + \sum_{i=1}^{n_c} k_{i,j} + \sum_{i=1}^{n_c} [G_1(1/N, 1/\lambda) - G_2(1/N, 1/\lambda)] \\ &= \sum_{i=1}^{n_c} X_t^i + L(\lambda).\end{aligned}$$

   Here summation is over all pairs of keys, which collectively sum to zero. The Gamma difference noise sums to Laplace noise in the aggregate cluster sum [8, 1].

$\hat{S}_t$ for DREAM is an unbiased estimator for S_t. The Laplace noise has mean zero, meaning the expected value of $\hat{S}_t$ is S_t with λ fixed across each cluster at each time point. In addition, an advantage of DREAM is the elimination of the middle-man aggregator, or alternatively, trust of the electricity supplier. The encryption scheme ensures that no agent of the electricity suppliers, authorised or not, can access individual meters' time series. It also protects from the privacy risk of transmission, since intercepting agents would not be able to decrypt any signals unless in possession of signals from all households.

3.2 Clustering

DREAM perform analysis on the clusters of smart meters [1]. In many cases, electricity suppliers will have contextual guidance in how to form the clusters. In order to give informative insights, it is often required to group the clusters based on proximity in the electricity grid.

However, if electricity suppliers have no natural way of grouping the meters, then it stands to ask, is there a methodology of clustering that has advantage to the masking? Recall that the net effect of the DREAM procedure is privately produced noisy aggregates, using Laplacian noise calibrated for ϵ-differential privacy. Specifically, the dispersion pa-

rameter of the noise is

$$\lambda_\epsilon(\mathbf{X}) = \frac{\max_i\{X_t^i\}}{\epsilon}.$$

This parameter must be set across the entire cluster, requiring communication between nodes in the cluster to find the maximum. In addition, in order to form an upper bound sufficient for differential privacy, the definition of $\lambda_\epsilon(\mathbf{X})$ using the maximum observation is necessarily highly sensitive to the extreme values of the data.

The authors proposed an approach of choosing clusters, aptly dubbed *smart clustering*. Suppose there were a way for the electricity supplier to estimate the average usage level of all the households in a geographical area, such as by ranking quarterly bills for each household. This would allow clustering with greater homogeneity, lowering the unnecessary error introduced into the noisy sums. For example, when there is no geographical context given, the smart meters are sorted according to their daily averages over the period for which data is available. From this sorted list, the clusters are formed by consecutive groups of n_c, with the last group containing n_c plus the extra leftovers. Here n_c is a sufficiently large number avoiding a risk of forming an small aggregate.

We conducted simulations to compare the performance smart clustering and random allocation. Smart clustering produces a more efficient data protection scheme than random cluster allocation: the same degree of differential privacy is achieved, but at a lower cost of relative error for subsequent analysis performed by the electricity supplier or other data users. The simulation results is available upon request.

Note that DREAM has a major practical drawback. The decryption process requires the encrypted signals from each and every node to succeed. It is possible to have one or more nodes failing to communicate their measurements at a particular time point, due to blackout, circuit breaker activation, failure in telecommunications or failure of the node hardware itself. In this case, decryption is no longer available and so the sum estimation at the time point becomes impossible. Granted, if the unresponsive node(s) is able to store its encrypted value for retransmission at a later date, this vulnerability is reduced. A solution is given in the next section.

4 The Multiplicative Noise Method

One way to tackle the missing data issue discussed in the previous section is to implement multiplicative noise to household smart meter data. Noise multiplication has a long history of being used for data masking, mainly in official statistics and remote sensing imagery (see Evans [4], Evans et al. [5], Kim and Winkler [7], Corner et al. [3] and others). Nayak et al. [12] reviewed the additive noise method and the multiplicative noise method, and they pointed out that it is difficult to generate the noise values under the additive noise method given that the noise distribution depends on the original data values. By contrast the multiplicative noise method is able to provide uniform record level protection in terms of noise coefficient of variation to all values.

In this article, we propose a data-masking method using a multiplicative technique for smart meter data, and discuss distributions appropriate for such noise to result in the greatest protective benefit. The difference between the multiplicative scheme and the DREAM scheme is that the multiplicative scheme involves transmitting perturbed individual values from each node to the electricity supplier without encryption. In addition to this simplification, the scheme is robust to failure of nodes, since the absence of any values does not threaten the estimation of the aggregate sum, unlike the DREAM scheme.

In a setting similar to the DREAM, we consider an electricity supplier desiring the aggregate sum of consumption from a group of n_c households' smart meters at each time t. The smart meter at each house records its consumption levels as a regularly-spaced time-series, with the measurement from the ith node at time t denoted X_t^i. Our goal is again to ensure that the electricity supplier is able to estimate the true cluster sum $S_t^c = \sum_{i=1}^{n_c} X_t^i$ without disclosure of the individual household values $\{X_t^i\}$, in order to protect the privacy of each household.

In general, the concept of employing Multiplicative Noise is to simply calculate the perturbed values $\tilde{X}_t$, such that

$$\tilde{X}_t = X_t \times M_t,$$

where M_t is a random variable sampled independently from a distribution with mean μ_M and standard deviation $\text{sd}(M)$ for time t.

The electricity supplier calculates the noisy sum for all households, $\tilde{S}_t^c = \sum_{i=1}^{n_c} \tilde{X}_t^i = \sum_{i=1}^{n_c} X_t^i M_t^i$ and subsequently, $\hat{S}_{t,M}^c = \tilde{S}_t^c / \mu_M$ which satisfies the following properties:

Property 1. Estimator $\hat{S}_t^c$ is an unbiased estimator for the true sum S_t^c given the true values $\{X_t^i\}$;

Property 2. The conditional standard error of the estimated sum $\hat{S}_{t,M}^c$ is $\left\{ sd(M)/\mu_M \right\} \sqrt{\sum_{i=1}^{n_c} \left(X_t^i\right)^2}$, given $\{X_t^i\}$.

Clearly this multiplicative scheme is simple in its formulation, but care must be given to pick a distribution for M with advantageous characteristics for masking. Unlike the DREAM scheme of the previous section, the distribution of the noise must be known to the electricity supplier–or at the very least, its mean–in order to use μ_M for the estimate $\hat{S}_{t,M}$.

4.1 Noise Distribution: Twin Uniform Distribution

Given that both perturbed values and the mean of the noise often are known to the electricity supplier, the distribution of the multiplicative noise plays an important role in protecting the values of the original data. An obvious choice would be a simple uniform distribution. The problem with the simple uniform distribution is that with a desired level of accuracy in estimation of the sum, the range of the parameters (maximum noise - minimum noise) need to be small, and as a result the distribution allows too much central tendency in the noise M, i.e., the masked data is not sufficiently perturbed from the original values. Broadly speaking, we need a distribution with less central tendency. The first and most obvious suggestion is to simply increase the dispersion of M, i.e., increase the maximum error parameter, but this necessarily reduces the accuracy of the electricity supplier's perturbed sum, $\hat{S}_{t,M}$.

Instead, we consider a multiplicative noise M with *twin uniform distribution*, TwinUnif($\mu, \alpha_{\min}, \alpha_{\max}$), which has the following density,

$$f_M(m) = \begin{cases} \frac{1}{2\mu(\alpha_{\max} - \alpha_{\min})}, & m \in [\mu(1-\alpha_{\max}), \mu(1-\alpha_{\min})] \cup [\mu(1+\alpha_{\min}), \mu(1+\alpha_{\max})] \\ 0, & \text{else} \end{cases}$$

As Figure 1 shows, the twin uniform distribution is essentially a mixture distribution of the two uniform distributions Unif($\mu(1-\alpha_{\max}), \mu(1-\alpha_{\min})$) and Unif($\mu(1+\alpha_{\min}), \mu(1+$

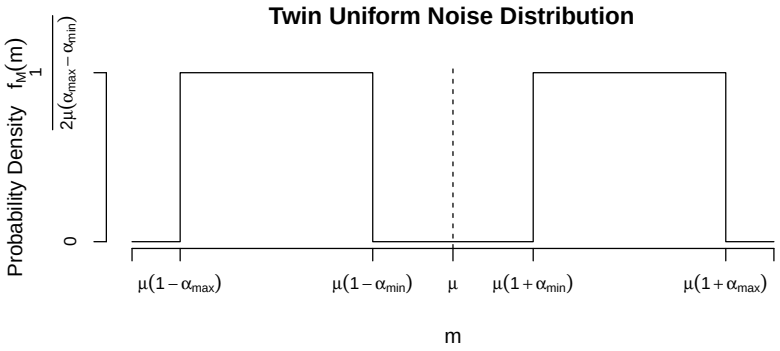


Figure 1: Twin uniform distribution, showing the gap around the mean μ that prevents accurate estimation of X_t^i by the electricity supplier.

$\alpha_{\max}$)), mixed with equal weight of probability. It can also be written as $M \overset{d}{=} \mu(1 + S \times C)$, where $\mu > 0$ is the arbitrary scaling constant, S ('sign') takes values in $\{-1, 1\}$ with equal probability, and $C \sim \text{Unif}(\alpha_{\min}, \alpha_{max})$ independently of S. The twin uniform distribution is chosen for simulation in this paper, though other distributions are also possible. The distribution is relatively simple, making it feasible for implementation in the inexpensive hardware of typical smart meters. We have

$$E[M] = \mu$$

and

$$\text{sd}(M) = \frac{\mu\sqrt{\alpha_{\max}^2 + \alpha_{\max}\alpha_{\min} + \alpha_{\min}^2}}{\sqrt{3}}\,. \tag{1}$$

Following Property 2., the conditional standard error of the sum estimator for a particular cluster of size n_c with a twin uniform distribution noise is

$$\text{sd}(\hat{S}_{t,M}^c) = \frac{\sqrt{\alpha_{\max}^2 + \alpha_{\max}\alpha_{\min} + \alpha_{\min}^2}}{\sqrt{3}} \sqrt{\sum_{i=1}^{n_c} \left(X_t^i\right)^2}\,.$$

Consider the estimate $\hat{X}_{t,M} = \tilde{X}_t/\mu = X_t M_t/\mu$, calculated using only the perturbed value and the mean of the noise, both of which are known to the electricity supplier. We term $\hat{X}_{t,M}$ the *central estimator*, as it uses the central mean μ of the Twin Uniform distribution. $\hat{X}_{t,M}$ has conditional expectation $E[\hat{X}_{t,M}^i | X_t^i] = X_t^i(E[M]/\mu) = X_t^i$, so $\hat{X}_{t,M}$ is an unbiased estimate for the original values. The disclosure risk for individual households for multiplicative method with twin uniform noise, using the probability of estimate relative error below a threshold δ is

$$\begin{aligned} p_{\delta,X}(\alpha_{\max}, \alpha_{\min}) &= P\left(\left|\frac{\hat{X}_{t,M} - X_t}{X_t}\right| < \delta\right) = P\left(\left|\frac{M}{\mu} - 1\right| < \delta\right) \\ &= P\{\mu(1-\delta) < M < \mu(1+\delta)\}\,. \end{aligned}$$

Depending on the value of δ in relation to the distribution parameter, the analytic disclosure risk varies. When $\delta \leq \alpha_{\min}$ or $\delta \geq \alpha_{\max}$, the disclosure risk is 0 and 1, respectively. When $\alpha_{\min} < \delta < \alpha_{\max}$, the disclosure risk is

$$p_{\delta,X}(\alpha_{\max}, \alpha_{\min}) = \frac{\delta - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}\,.$$

4.2 Secondary Disclosure Risks

A further consideration of the Multiplicative Noise masking scheme is the issue of regression risk. As there is no encryption applied to the values sent to the electricity suppliers, all received values may be leveraged by the electricity suppliers for better malicious estimation than afforded by the central estimator $\hat{X}_{t,M}$, by applying a linear regression model to the received values $\{\tilde{X}_t^i\}$. This presents an additional form of privacy risk which must be managed by adequate noise levels, i.e. sufficiently high $\alpha_{\max}$ given a defined $\alpha_{\min}$.

Moreover, the electricity suppliers could choose to take an alternative approach to estimation the unmasked values. Rather than producing one estimate $\hat{X}_{t,M}$ using the expected value of the entire distribution M, they could instead produce two estimators using the expected values of each component uniform distribution in the mixture M. The *lower and upper estimators* are

$$\hat{X}^1_{t,M} = \frac{\tilde{X}_t}{\mu_{M;1}} = \frac{X_t M_t}{\frac{1}{2}\mu(1 - \alpha_{\max} + 1 - \alpha_{\min})} = \frac{X_t M_t}{\mu(1 - \frac{\alpha_{\max}+\alpha_{\min}}{2})},$$

$$\hat{X}^2_{t,M} = \frac{\tilde{X}_t}{\mu_{M;2}} = \frac{X_t M_t}{\frac{1}{2}\mu(1 + \alpha_{\max} + 1 + \alpha_{\min})} = \frac{X_t M_t}{\mu(1 + \frac{\alpha_{\max}+\alpha_{\min}}{2})}.$$

The electricity suppliers must also make a guess at which estimator to use, the lower or upper. In many situations this guess could be educated by historical trends, lending more credibility to one estimate over the other, but there is still no certainty to such guesswork.

The main risk of this type of estimation is that it nullifies the protection garnered by the central gap in the support of the twin uniform distribution, as parametrised by $\alpha_{\min}$. The scaling factors $\mu_{M;1}$ and $\mu_{M;2}$ are simply the midpoints of each component uniform distribution (refer to Figure 1), and so the efficiency of each estimator over its relevant support is entirely dependent on the width of that support, the size of the interval $(\alpha_{\min}, \alpha_{\max})$.

As an extreme limiting case, if the interval is reduced to nothing, i.e., $\alpha_{\max} \to \alpha_{\min}$, then the twin uniform noise distribution for M degenerates into a discrete distribution, with probability 0.5 gathered at the points $\mu(1 \pm \alpha_{\min})$. The effect for estimation in this limiting case is that one of the estimates, either $\hat{X}^1_{t,M}$ or $\hat{X}^2_{t,M}$, *perfectly* estimates the true value X_t.

This goes to show that, no matter the width of the $\alpha_{\min}$-controlled central gap, it is still necessary to ensure sufficiently high $\alpha_{\max}$, lest the estimators $\hat{X}^1_{t,M}$, $\hat{X}^2_{t,M}$ become too effective, or the correlation $\text{corr}(\hat{X}_{t,M}, X_t)$ becomes too high, both of which present other disclosure risks beyond simple estimation using the central estimator $\hat{X}_{t,M}$.

4.3 Shifting

The masking algorithm described above makes an implicit assumption: that $X_t^i > 0$ for the ith household at all time points t. Non-negativity alone is valid, given current should never

flow in the opposite direction to ordinary power usage. However, it is not impossible for a zero value to occur.

This creates two primary issues. First, from the practical perspective of the simulation metrics, the relative error $p_\delta(\alpha_{\max}, \alpha_{\min})$ does not exist due to division by zero. This itself is not insurmountable, but it is desirable to avoid it. The second issue is more serious: multiplicative noise applied to a signal $X_t^i = 0$ causes no perturbation, thus the masking is nullified. This poses a high disclosure risk for households with zero signals. Indeed, given that zero signals are often a result of dwelling vacancy that persists across many hours, even days, zero signal households must be protected.

Lin [10] offered a solution to the issues from zero signals by formulating a minor modification to the scheme, referred to as *shifting*. The application of the shifting approach can be found in Wakefield and Lin [13]. The algorithm for the Multiplicative Noise method is given as follows:

Algorithm: Multiplicative Noise with Shifting

1. Each node shifts their measurement by a value a, which is known to the electricity suppliers :
   $Y_t = X_t + a, \qquad a > 0.$
2. Each node samples from the distribution M, multiplies:
   $\tilde{Y}_t = Y_t \times M_t = (X_t + a)M_t.$
3. The electricity suppliers receives noisy, shifted values $\{\tilde{Y}_t^i\}$, aggregates to obtain $\tilde{S}_t^c = \sum_{i=1}^{n_c} \tilde{Y}_t^i = \sum_{i=1}^{n_c} (X_t^i + a)M_t^i$.
4. The electricity suppliers uses the noise mean μ_M to obtain the unbiased estimate for the cth cluster:

$$\hat{S}_{t,M}^c = \frac{\tilde{S}_t^c}{\mu_M} - n_c a = \frac{\sum_{i=1}^{n_c} \tilde{Y}_t^i}{\mu_M} - n_c a = \frac{\sum_{i=1}^{n_c} (X_t^i + a)M_t^i}{\mu_M} - n_c a\,.$$

The expression for the sum estimate $\hat{S}_t$ is only slightly different to previous; it not only de-scales as before, but then de-shifts the perturbed sum $\tilde{S}_t$. It is unbiased in the following sense:

$$\begin{aligned} E[\hat{S}_{t,M}^{n_c} | \{X_t^i\}] &= E\left[\frac{\sum_{i=1}^{n_c} (X_t^i + a)M_t^i}{\mu_M} - n_c a \middle| \{X_t^i\} \right] = \sum_{i=1}^{n_c} \left[(X_t^i + a) \frac{E[M_t^i]}{\mu_M} \right] - n_c a \\ &= \sum_{i=1}^{N} X_t^i + (n_c a - n_c a) = S_t^{n_c}\,. \end{aligned}$$

Here $E[M_t^i] = \mu_M$. The malicious estimate of the true household value X_t using $\tilde{Y}_t$ is

$$\hat{X}_{t,M} = \frac{\tilde{Y}_t}{\mu_M} - a = \frac{(X_t + a)M_t}{\mu_M} - a\,,$$

which is an unbiased estimator for X_t, i.e.,

$$E(\hat{X}_{t,M} | X_t) = E\left[\frac{(X_t + a)M_t}{\mu_M} - a \middle| X_t \right] = X_t\,.$$

For the privacy metric p_δ ,and correspondingly, the sample proportion $\hat{p}_\delta$, with shifting added to the scheme we now use the relative error of the central estimator $\hat{Y}_{t,M} = \tilde{Y}_t/\mu_M = (Y_t M_t)/\mu_M$ of Y_t, rather than $\hat{X}_{t,M}$ of X_t. We are essentially comparing shifted estimate against shifted true value, and this avoids division by zero as described above.

To the attentive reader, this may raise the question: '*So what are we really protecting now, the true value X_t or the shifted true value Y_t?*' It is true that since we apply an additive transformation to form Y_t, the relative error on the scale of Y_t is not equal to the relative error on the scale of X_t. The argument we make as resolution is that if the values Y_t are sufficiently protected (i.e. $p_{\delta,Y}$ using $\hat{Y}_{t,M}$ values is sufficiently low), then X_t^i is sufficiently protected also. Conversely, if Y_t is poorly protected and vulnerable to disclosure, i.e. $p_{\delta,Y}$ is high, then X_t is also vulnerable, since it can be estimated accurately by a simple shift of an accurate estimate $\hat{Y}_{t,M}$.

5 Comparison Study

In this section, we compare two data masking methods, DREAM and Multiplicative Noise, using the ESSNET data described in Section 2. We set the cluster size of $n_c = 100$ for $N = 3371$ households, and the results are calculated for 96 hourly time points (4 days). For both schemes, sums are computed over the same clusters, which are decided using the Smart clustering algorithm as described in Section 3.2.

Here we discuss the parameters defined in each data masking method. DREAM scheme has one parameter ϵ. We consider various values for ϵ, ranging from 0.025 to 2. For reference $\epsilon = 1$ is considered as a moderate level of differential privacy by Ács and Castelluccia [1]. For the Multiplicative Noise scheme, we consider $\mu = 27$, $a = 0.6$, $\alpha_{\min} = 0.1$ and various levels of $\alpha_{\max}$. Of these levels of $\alpha_{\max}$, $\alpha_{\max} = 0.5$ was chosen, in combination with shift value $a = 0.6$ to satisfy the correlation $\text{corr}(\hat{Y}_{t,M}, Y_t) \leq 0.8$, although the desired level of the correlation constraint on regression risk is arbitrary. A general rule of thumb is that the smaller correlation is, the better the privacy [11].

The following of this section compares the performance of the two data masking methods in the following three aspects: *privacy*, in minimising risk of disclosure of the true household signal X_t; *utility*, in maximising accuracy of estimation of the sum over a cluster S_t; and *practicality*, in considering the technical feasibility, reliability and costs of each scheme. We use this study to indicate that using the Multiplicative Noise method to protect data can have similar outcome to using the DREAM. More importantly, the approach of the Multiplicative Noise method is more robust than the approach of DREAM.

5.1 Privacy

To begin with, DREAM both directly and indirectly protects the privacy of individual households. Through the addition of a randomly-generated encryption key, each perturbed signal $\tilde{X}_t$ is unable to be determined individually. Moreover, the Gamma difference noise on each individual signal aggregates to form Laplace noise in the sum $\hat{S}_t$, which ensures ϵ-differential privacy. This differential privacy is important for protection against malicious extraction techniques that make use of external data sources.

On the contrary, the Multiplicative scheme offers far less assurance of privacy. It does not satisfy any formal standard, such as differential privacy. Instead we firstly measure disclosure risk using the *probability of estimation within a relative error threshold δ*, where we

consider only the central estimator $\hat{Y}_{t,M}$. This is estimated by the sample proportion $\hat{p}_{\delta,Y}$.

$$p_{\delta,Y}(t) = P\left(\left|\frac{\hat{Y}_{t,M} - Y_t}{Y_t}\right| < \delta\right), \qquad \hat{p}_{\delta,Y}(t) = \frac{1}{N}\sum_{i=1}^{N} I\left(\left|\frac{\hat{y}^i_{t,M} - y^i_t}{y^i_t}\right| < \delta\right)$$

We choose $\alpha_{min} = \delta = 0.1$ and $p_{\delta,Y}(t) = 0$ for all t.

Secondary disclosure risks of the Multiplicative Noise scheme arise in considering other forms of estimation by malicious agents. We consider *regression risk*, the risk of improved estimation by fitting a model to the received data $\tilde{Y}_t$. This is measured by the correlation of the central estimator with its true values, corr$(\hat{Y}_{t,M}, Y_t)$. Figure 2 displays the sample correlations across time. A large α_{max} ($\alpha_{max} \geq 0.4$) limits the correlations approximately to 0.8.

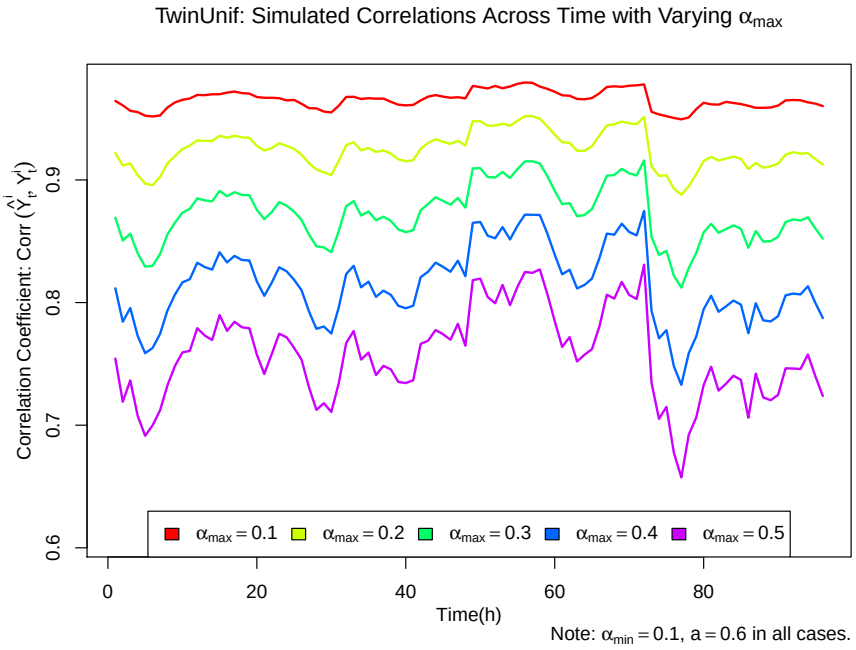


Figure 2: Correlations for the Multiplicative scheme, corr$(\hat{Y}_{t,M}, Y_t)$.

Though not measured directly, another risk arises from estimation using the means of each component uniform distribution of the Twin Uniform, producing the lower and upper estimators $\hat{X}^1_{t,M}$ and $\hat{X}^2_{t,M}$. We conjecture that reduction of regression risk also reduces the risk posed by the lower and upper estimators. Further research is required to investigate this conjecture.

In summary, where DREAM faces little to no privacy issues under normal parameter values, the Multiplicative scheme must be calibrated heavily in order to have any privacy assurance. In this manner, DREAM far surpasses the Multiplicative scheme in protecting the signals of smart meters.

5.2 Utility

DREAM and the multiplicative scheme share the goal of producing accurate estimates for the sum $S_t^c = \sum_{i=1}^{n_c} X_t^i$ over each cluster of dwellings. Unlike privacy, we can use the same utility measures for both schemes, allowing direct comparison.

Recall that in the previous sections, we used $\hat{S}_t^c$ and $\hat{S}_{t,M}^c$ for the estimates of S_t^c in DREAM and the Multiplicative Noise scheme, respectively. For convenience, we used the notation $\hat{S}_t^c$ for the estimate of S_t^c for both schemes in this section.

The first measure is the *probability of sum estimation within a relative error threshold* δ, $p_{\delta,S}$. As before in measuring privacy, this is estimated using the sample proportion $\hat{p}_{\delta,S}$.

$$p_{\delta,S} = P\left(\left|\frac{\hat{S}_t - S_t}{S_t}\right| < \delta\right), \qquad \hat{p}_{\delta,S} = \frac{1}{C}\sum_{c=1}^{C} I\left(\left|\frac{\hat{s}_t^c - s_t^c}{s_t^c}\right| < \delta\right)$$

Unlike the application of $p_{\delta,Y}$ to the privacy of Y_t, for S_t, we wish $p_{\delta,S}$ to be as close to 1 as possible.

Figure 3 shows the accuracy rates $\hat{p}_{\delta,S}$ of S_t, with accuracy threshold $\delta = 0.1$. Taking careful note of the different vertical scales of the two plots, the plots demonstrate that Multiplicative Noise *can* achieve accuracy rates comparable to DREAM, such as $\alpha_{\max} = 0.2$ compared with $\epsilon = 1$. However, this comes at the sacrifice of privacy; $\alpha_{\max} = 0.2$ was shown to be insufficient to ensure regression risk below tolerance 0.8, and the compliant level $\alpha_{\max} = 0.5$ has far inferior error rates to the standard DREAM setting of $\epsilon = 1$. In short, the Multiplicative Noise scheme can achieve similar accuracy rates to DREAM, but not whilst satisfying the correlation constraint and ensuring reasonable privacy.

The next measures of sum accuracy are the *mean relative error* (MRE) and *mean unsigned relative error* (MURE),

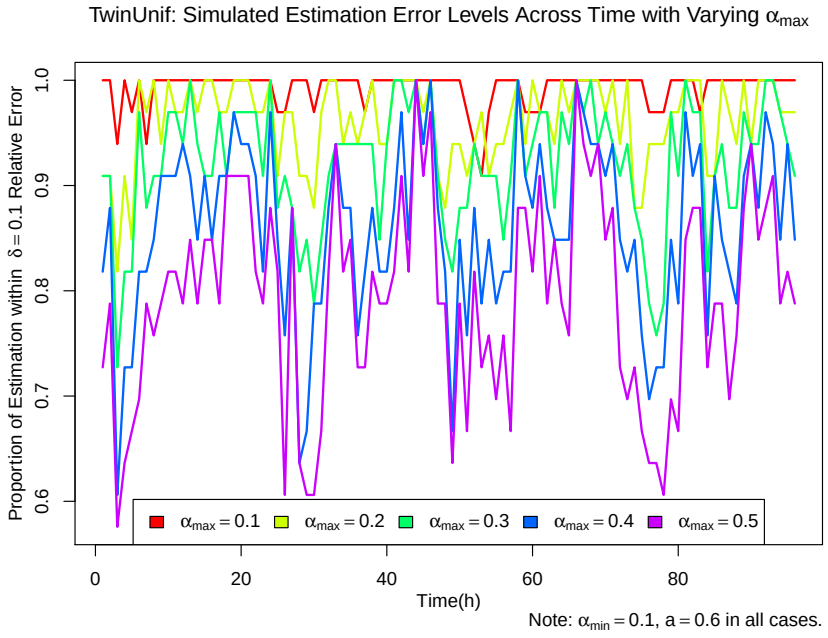
$$\text{MRE} = \frac{1}{C}\sum_{c=1}^{C}\frac{\hat{s}_t^c - s_t^c}{s_t^c}, \qquad \text{MURE} = \frac{1}{C}\sum_{c=1}^{C}\left|\frac{\hat{s}_t^c - s_t^c}{s_t^c}\right|.$$

where the sums are taken over all clusters $c = 1, \ldots, C$ at each time t. As explained in Section 2, distinct from the rate of error $p_{\delta,S}$, MRE indicates the overall or average relative error. Since this includes both overestimation and underestimation, the errors of opposite sign offset each other, obscuring the magnitude of the relative errors themselves but giving an idea of the overall bias at each time point. MURE eliminates sign and the accompanying offset, providing a better measure of the average magnitude of relative error.

Figure 4 displays the MRE for both schemes across time. In the first plot, for $\epsilon > 1$ (standard privacy or weaker) there appears no MRE beyond $\pm 1.5\%$ at any time point, though lower ϵ values display much greater deviance. In the second plot, similar can be said for $\alpha_{\max} = 0.1, 0.2$, but not for the privacy compliant $\alpha_{\max} = 0.5$, which exhibits much larger deviations from zero. On the positive side, both schemes display no consistent bias, being centred at zero overall in both cases. This comes as no surprise given the sum estimators of each scheme are unbiased.

Figure 5 displays the simulated MURE for both schemes across time. Taking careful note of the differing vertical scales once again, we see that the Multiplicative scheme is capable of MURE on par with DREAM, with the $\alpha_{\max} = 0.2$ MURE curve in a similar range around 3% as $\epsilon = 1$. However, the curve with acceptable privacy assurance, i.e., $\alpha_{\max} = 0.5$, has MURE averaging around 6-8%, reaching as high as 10-12%. These averages are disappointing enough on their own, not to mention that necessarily some individual values will be in far excess of these averages.

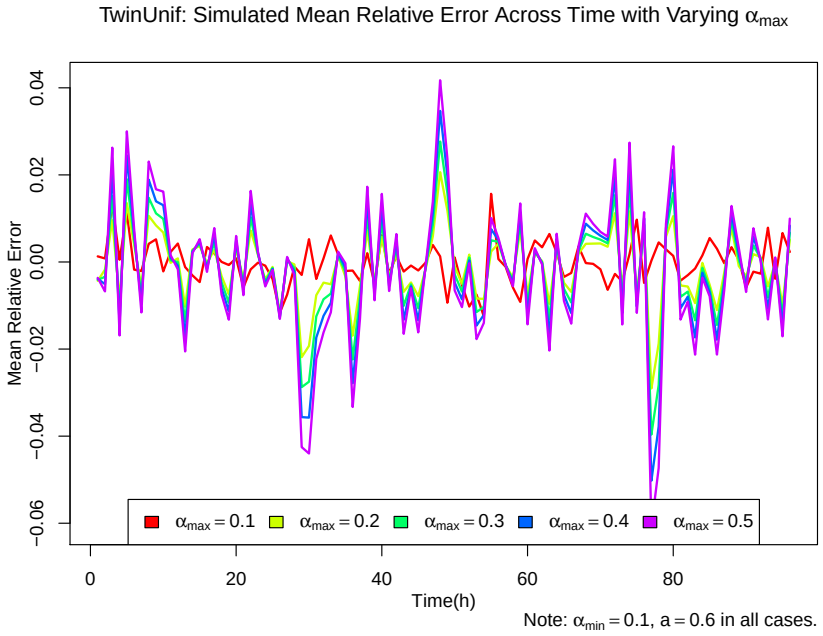
(a) DREAM



(b) Multiplicative Noise (Twin Uniform)

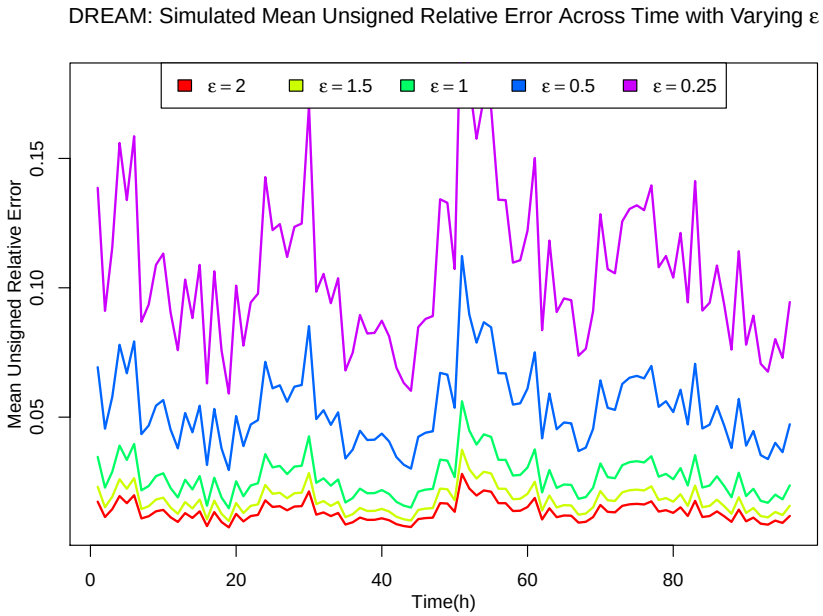
Figure 3: Simulated sum accuracy rates $\hat{p}_\delta$ with accuracy threshold $\delta = 0.1$.

(a) DREAM

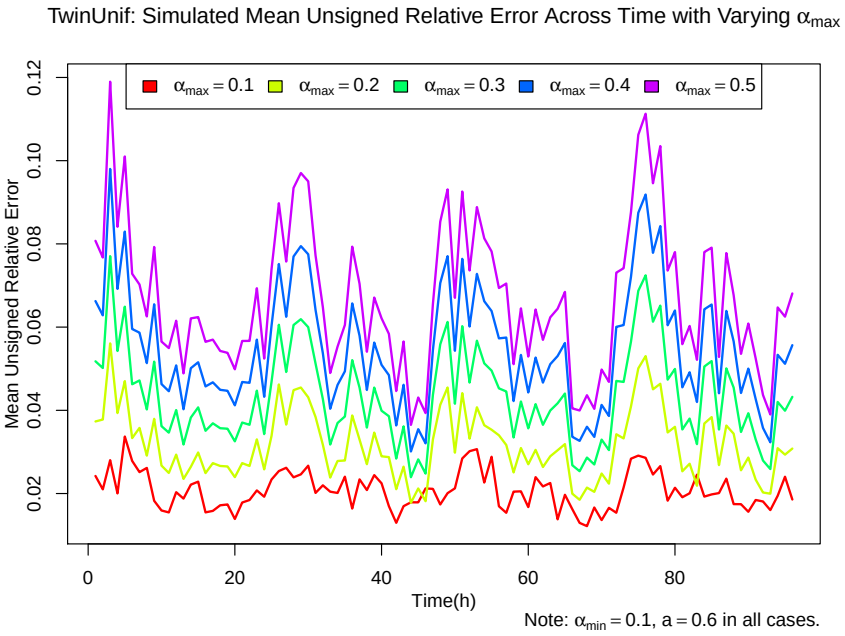


(b) Multiplicative Noise (Twin Uniform)

Figure 4: Simulated mean relative error (MRE) of the two masking schemes across time.



(a) DREAM



(b) Multiplicative Noise (Twin Uniform)

Figure 5: Simulated mean unsigned relative error (MURE) of the two masking schemes across time.

In summary, DREAM possesses advantage in sum estimation accuracy measures to the Multiplicative scheme due to the latter's restrictive parameter constraints to ensure some modicum of data privacy. With careful calibration of the masking parameters, the Multiplicative scheme can achieve more accurate sum estimation in line with DREAM.

5.3 Practicality

Beyond the numerical properties of privacy and utility, there are other characteristics of each scheme to factor into evaluation.

DREAM requires intercommunication between nodes, both to acquire the maximum measurement in the cluster at each time point in computing $\lambda = \max_i\{X_t^i\}/\epsilon$, and in establishing the randomly-generated private encryption keys. These functions, and the capacity for the nodes to privately initiate a peer-to-peer network independent of the electricity supplier's assistance, require a certain level of hardware sophistication. Such hardware comes with additional cost, which will ultimately need to be absorbed as an additional expense of the energy provider (perhaps offset against the revenue produced by insights driven by smart meter data), given very few consumers are receptive of outlay for which they have no personal benefit.

By contrast, the required node capability of the Multiplicative scheme is significantly lesser. Peer-to-peer networking is not necessary, only the simpler bidirectional communication with the energy provider's central information systems. The node is still required to respond to parameter adjustments, draw random samples from the distribution M and multiply values as appropriate, but these operations are well within the capabilities of a simple microprocessor. This presents a lower outlay for node installation from the energy provider.

Additionally, the Multiplicative scheme has superior robustness to technical failure. The absence of a small number of nodes from a cluster masked under the Multiplicative scheme will cause underestimation of S_t, but this is not catastrophic. The electricity suppliers can still form a reasonable makeshift estimate by simply scaling up the received noisy sum $\tilde{S}_t$ by a factor $n_c/(n_c - f)$, where n_c denotes the number of nodes in the cluster, f the number of 'failed' or absent nodes.

The absence of a small number of nodes' measurements from DREAM has disastrous consequences. Though privacy is not compromised, sum estimation is impossible in the absence of even one node's measurement, since the cluster cannot be decrypted without the participation of all nodes. To put this into perspective, where the Multiplicative scheme may have frequently high error in its sum estimation, this could be considered less serious than the threat of no possible estimation, error-ridden or not, for a particular time point.

Finally, the Multiplicative scheme is more flexible in its choice of clustering. Since each individual node's measurements are sent quite independently of other nodes, clusters may be formed and reformed using these different combinations of dwellings decided after the fact. Any sufficiently large cluster will have comparable probability of accuracy as any other large cluster.

Conversely, DREAM requires clusters to be set in advance to enable the computation of the parameter λ and encryption, producing a single sum estimate as output after aggregation and decryption. The encrypted values cannot be used for any other cluster of dwellings than as they were produced. Furthermore, the electricity suppliers cannot be allowed to request the nodes to form different clusters after the fact and resend encrypted values in this new grouping, as this would constitute a very serious compromise to the integrity of

the scheme. In this way, DREAM does not provide nearly so much analytical flexibility as the Multiplicative scheme.

In summary, DREAM's superior numerical properties come attached with increased financial cost, complexity of hardware and network infrastructure, decreased robustness to technical failure and decreased analytical flexibility, compared to the Multiplicative scheme.

6 Concluding Remarks

6.1 Conclusion

Taking into account only the first two criteria of evaluation, numerical measures of privacy and utility, the DREAM scheme is superior in achieving a more efficient trade-off of privacy for accurate sum estimation (and vice versa). The Multiplicative scheme has a less efficient trade-off, with the simultaneous achievement of adequate privacy and adequately accurate sum estimation through parameter calibration difficult. Also, the transmission of unencrypted values to the electricity suppliers presents a number of secondary disclosure risks such as regression and alternative malicious estimation methods - risks that are avoided by the encryption in DREAM. For these reasons, the Multiplicative Noise is currently not a viable alternative to DREAM without careful parameter selection.

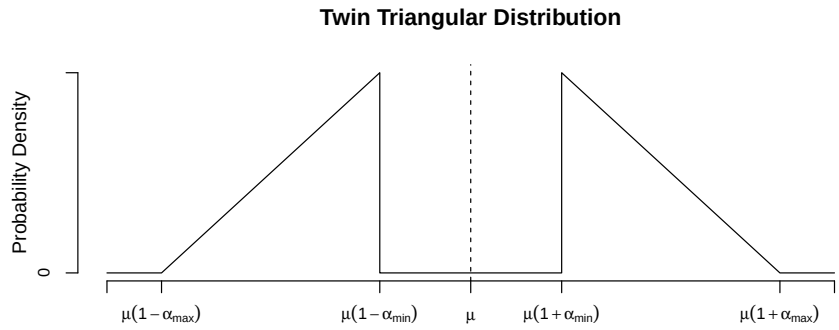
Factoring in practical aspects, the Multiplicative scheme becomes a more attractive option than previously. In the areas of cost, reliability and flexibility, the Multiplicative scheme is superior to DREAM. If future research is able to achieve a better trade-off of privacy and utility, the Multiplicative scheme may overtake DREAM as the preferred scheme due to its economic and practical advantages.

6.2 Future Work

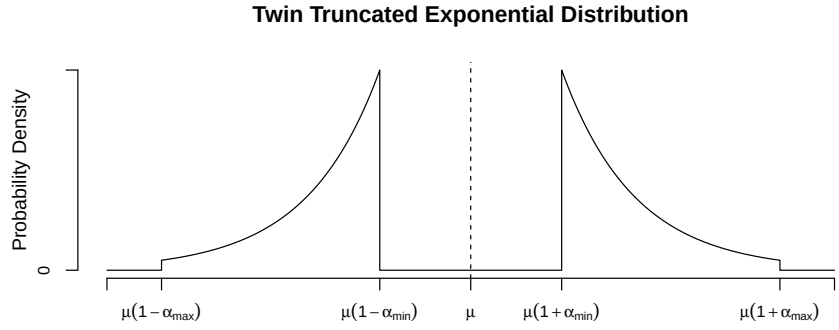
As discussed above, the main flaw in the Multiplicative scheme's viability as a competitor of DREAM is that it is tricky to achieve an efficient trade-off of privacy for accurate sum estimation. The Multiplicative algorithm in full generality has great practical merit, but investigation into the noise distribution M is much needed.

The valuable feature of the Twin Uniform distribution is its simplicity and the capacity to limit the accuracy of the central estimate $\hat{Y}_t$ through the gap in support about its mean. Such a gap can be created using other mixture distributions beyond the uniform distribution. Figure 6 provides two possible distributions, the Twin Triangular distribution and the Twin (Truncated) Exponential distribution. Both the Twin Triangular and Twin Truncated Exponential distributions concentrate likelihood more heavily towards the mean of the distribution than the Twin Uniform. This ought to reduce error overall, though it may raise the correlation to an unacceptable level of regression risk.

Another idea of extension for the Multiplicative scheme is to consider the values chosen for the masking parameters. Given there are multiple masking parameters in the Multiplicative scheme, an optimisation process can be proposed in order to maximize accuracy and privacy simultaneously. In addition, parameters are not required to be consistent across the different clusters, with only requirement being a consistent mean μ_M across the family of distributions. By allowing variability of masking parameters between and within clusters, it is plausible to achieve greater accuracy. On an absolute scale of error, larger signals within a cluster are more influential in deviating the sum estimate $\hat{S}_t$ from its true value. For instance, +40% relative error on a signal of 0.3KWh has far less impact than +40%



(a) Twin Triangular Distribution



(b) Twin Truncated Exponential / Split Laplace Distribution

Figure 6: Potential future distributions.

relative error on a signal of 3KWh. It seems it would be more efficient, then, to allow greater error in households of small average magnitude and lesser error in households of large average magnitude. This could be achieved by setting different values of $(\alpha_{\max}, \alpha_{\min}, a)$ for different households.

References

[1] G. Ács and C. Castelluccia. I have a dream!(differentially private smart metering). *Information hiding*, 6958:118–132, 2011.

[2] J.M. Bohli, C. Sorge, and O. Ugus. A privacy model for smart metering. *Communications Workshops (ICC), 2010 IEEE International Conference*, 1:1–5, 2010.

[3] B. R. Corner, R. M. Narayanan, and S. E. Reichenbach. Noise estimation in remote sensing imagery using data masking. *International Journal of Romote Sensing*, 24(4): 689–702, 2003.

[4] B. T. Evans. Effects on trend statistics of the use of multiplicative noise for disclosure limitation. *Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association*, 1997.

[5] T. Evans, L. Zayatz, and J. Slanta. Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistic*, 14(4):537–551, 1998.

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

[7] J. J. Kim and W. E. Winkler. Multiplicative noise for masking continuous data. *Proceedings of the Annual Meeting of the American Statistical Association*, 5(9):1–17, 2001.

[8] S. Kotz, T. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

[9] F. Laforet, E. Buchmann, and K. Böhm. Individual privacy constraints on time-series data. *Information Systems*, 54:74–91, 2015.

[10] Y. X. Lin. Protecting values close to zero under the multiplicative noise method. *UNESCO Chair in Data Privacy. Lecture Notes in Computer Sciences series 11126*, 1:247–262, 2018.

[11] Y. Ma, Y.X. Lin, and R. Sarathy. The vudream2011lnerability of multiplicative noise protection to correlation- attacks on continuous microdata. *Working Paper*, 2017.

[12] T. K. Nayak, B. Sinha, and L. Zayatz. Statistical properties of multiplicative noise masking for confidentiality protection. *Journal of Official Statistics*, 27(3):527–544, 2011.

[13] B. Wakefield and Y. X. Lin. Efficiency and sample size determination of protected data. *Privacy in Statistical Databases. UNESCO Chair in Data Privacy International Conference Proceddings*, 1:263–278, 2018.