IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 17, 2022, accepted February 27, 2022, date of publication March 8, 2022, date of current version March 16, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157390

Evaluation of Noise Distributions for Additive and Multiplicative Smart Meter Data Obfuscation

AHMED S. KHWAJA[1], (Senior Member, IEEE), SERHAT ERKUCUK[2], (Senior Member, IEEE), ALAGAN ANPALAGAN[1], (Senior Member, IEEE), AND BALA VENKATESH[1], (Senior Member, IEEE)

[1]Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada
[2]Department of Electrical-Electronics Engineering, Kadir Has University, 34083 Istanbul, Turkey

Corresponding author: Ahmed S. Khwaja (akhwaja@ryerson.ca)

ABSTRACT In this paper, we compare and analyze light-weight approaches for instantaneous smart meter (SM) data obfuscation from a group of consumers. In the literature, the common approach is to use additive Gaussian noise based SM data obfuscation. In order to investigate the effects of different approaches, we consider Gaussian, Rayleigh, generalized Gaussian and chi-square distributions to achieve either additive or multiplicative data obfuscation. For each type of obfuscation approach, we calculate the required parameters to achieve obfuscation such that 50% of the obfuscated data fall outside an interval equalling twice the mean of the instantaneous SM measurements. We also calculate the minimum number of SMs required to estimate the mean of the actual SM measurements, such that the estimate varies within only 0.5% of the actual mean with a 99.5% probability. Simulation results are used to verify the calculations, and it is shown that multiplicative Rayleigh and generalized Gaussian noise require the least number of SMs, which is 90% less than the traditional approach of additive Gaussian noise-based SM data obfuscation.

INDEX TERMS Smart meter, data obfuscation, additive noise, multiplicative noise.

I. INTRODUCTION

The emergence of smart grid has enabled two-way communication of electricity and data between suppliers and consumers. This communication is achieved based on the integration of information and communication technology with the energy and distribution infrastructures [1]. Smart meter (SM) is an important component of smart grids. The SMs are installed at customers' premises and provide real-time information about the energy usage [1] by transmitting their energy usage to a utility center [2].

This information can enable grid operators, distributors and suppliers to achieve electricity demand forecasting, fault detection, optimization of services [2], load balancing, dynamic pricing [3], etc. These benefits are achieved usually at the cost of privacy risks, which include leakage of information that can identify personal and behavioral information [3]. Using the periodic information transmitted by the SMs, an eavesdropper can gain information about the energy consumption behavior and habits of customers, as well as their presence or absence in their homes [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Yang Li.

Thus, it can be gathered that there are two conflicting issues that obstruct the use of SMs: 1) transmission of data from the consumers at periodic intervals so that the suppliers can utilize the data, 2) hiding the periodic data from eavesdroppers and possibly the suppliers to avoid invading the privacy of the consumers. To solve these issues, different privacy-preserving approaches have been proposed to protect the privacy while simultaneously ensuring extraction of useful information from the SM data. In the following paragraphs, we carry out a review of these approaches.

In [5], a rechargeable battery was used to modify meter readings, which could mask the actual energy consumption. In [6], the authors proposed an online control algorithm based on a Lyapunov function to simultaneously attain the objectives of protecting a consumer's privacy, while minimizing the electricity cost using a rechargeable battery. In [7], the author proposed the use of physically unclonable functions and channel state estimation to provide authentication and confidentiality, respectively for advanced metering infrastructure.

Pseudonymization provides the energy supplier (ES) with individual SM data, while the identity of each SM is kept hidden. This approach requires the presence of a trusted third

party. In [4], it was shown that an appearance of a new pseudonym or the disappearance of an existing pseudonym could be traced back to a new or departing customer, which could be used to identify a customer from the corresponding pseudonym. In [8], the authors showed that pseudonyms needed to be changed more frequently as the number of SMs increased.

In the aggregation approach, user privacy is protected by providing the ES with the aggregated power consumption of a group of SMs. The aggregated consumption is calculated by a trusted third party, or by using cryptographic methods [9]. Data encryption encodes the SM data using an encryption key. This approach is generally computationally intensive and requires coordination between different SMs and other entities. In [10] and [11], the authors used secure multiparty computation based on Shamir secret sharing algorithm and homomorphic encryption to protect user privacy. In [12], the authors presented a scheme to preserve privacy in communication between SMs and neighborhood gateways using hash and Exclusive-OR operations.

Unlike existing aggregating schemes that assumed only a single entity that receives the aggregated readings, the authors in [13] considered the presence of several entities that could access the aggregated data of different sets of SMs. Each SM used homomorphic encryption to send data to a local gateway. In [14], the authors used ElGamal encryption to protect against attacks on intermediate data collection nodes present between SMs and the ES. In [15], the authors proposed a privacy-preserving scheme where each SM homomorphically encrypted part of its data and sent them to other SMs. Subsequently, each SM then sent its aggregated readings to an aggregator.

The aforementioned privacy-preserving approaches are considered expensive, inefficient, poorly scalable, require SMs to communicate with one another, require a trusted third party, require generation and sharing of keys, etc. Another SM privacy-preserving approach is data obfuscation [16]–[25] that involves adding or multiplying the SM data with a random number. It is considered a light-weight approach that neither requires very extensive computations nor requires installation of separate infrastructure. The statistics such as periodic mean of a group of SM data can be estimated from the obfuscated data. However, there is a trade-off between privacy and utility: to achieve privacy for each SM user, a high variance noise should be added to the SM data, which in turn decreases the utility of the data. The mean or sum of the data should be estimated with a high accuracy, e.g., the authors in [26] provided an example for Brazil, where the relative percentage error for billing purposes must stay between $\pm 2\%$.

An existing SM data obfuscation approach based on additive noise requires a large number of consumers for accurate calculation of statistics from the obfuscated data [18]. While adding Gaussian noise is the most common approach for SM data obfuscation, consideration of multiplication and other distributions has not been pursued in general. In this paper, we address the lack of such approaches by evaluating the generation of random data using different distributions and considering both multiplicative and additive approaches.

We compare the results of Gaussian, Rayleigh, generalized Gaussian and chi-square distributions in terms of the noise variance required to achieve effective data obfuscation, and the minimum number of consumers needed to accurately calculate the mean value. To the best of our knowledge, this is the first time that these distributions have been proposed, compared and analyzed considering both additive and multiplicative approaches. This evaluation can serve as a reference in practical applications, providing merits or demerits of each data obfuscation approach. We further present guidelines for choosing a particular obfuscation approach. In the next section, we carry out a review of four types of obfuscation approaches: Additive, multiplicative, additive with other (additive^{+}), and miscellaneous approaches.

II. LITERATURE REVIEW

A. ADDITIVE

In [16], random data obfuscation was carried out using the Laplace distribution. In [17] and [18], random noise with Gaussian distribution was added to the consumers' energy usage data. These approaches could not accurately estimate the electricity consumption. In [23] and [24], the authors used additive noise to obfuscate SM data using Laplace distribution and uniformly distributed noise, respectively. In [27], the authors used correlated noise to achieve SM data obfuscation, and also proposed a generative adversarial networks (GANs)-based technique to achieve data obfuscation. The technique required the presence of a third party.

In [28], the authors used Laplace distribution and showed that by temporally smoothing the aggregated data, a group size of the order of thousands of SMs is sufficient for estimation of aggregated consumption. The authors in [29] introduced a model for local obfuscation of probability distribution by probability perturbation, which perturbed each single "point" datum by adding controlled probabilistic noise before sending it out to a data collector. The authors in [30] studied noise addition as a data privacy providing technique.

B. MULTIPLICATIVE

The authors in [31] proposed using truncated triangular distribution for masking survey data using multiplicative noise. In [32], the authors proposed data masking using orthogonal matrices, where the data were multiplied by orthogonal matrices before being released. The released data allowed exact statistics to be estimated from the masked data. In [33], the authors considered the case of using multiplicative noise to perturb original data. The authors showed that if the perturbed and original data were highly correlated, and if a malicious entity knew that the multiplicative noise was the data obfuscation mechanism, it could recover the original data by using linear regression for a large number of consumers.

The authors in [34] stated that multiplicative noise has the advantage of the size of perturbation being directly

TABLE 1. Comparison of existing data obfuscation techniques.

Ref. No.	Obfuscation technique	Smart grid	Comments
[16]	Additive	Yes	Used Laplace distribution
[17]	Additive	Yes	Used Gaussian distribution
[18]	Additive	Yes	Used Gaussian distribution
[23]	Additive	Yes	Used Laplace distribution
[24]	Additive	Yes	Used uniform distribution
[25]	Miscellaneous	Yes	Used GMM
[27]	Additive	Yes	Used correlated Gaussian noise
[28]	Additive	Yes	Used Laplace distribution
[29]	Additive	No	Used controlled probabilistic noise
[30]	Additive	No	Studied noise addition
[31]	Multiplicative	No	Used truncated triangular distribution
[32]	Multiplicative	No	Used orthogonal matrices
[33]	Multiplicative	No	Showed that if original and perturbed data are highly correlated, a malicious entity can recover the original data
[34]	Multiplicative	No	Stated that multiplicative noise has the advantage of perturbation size directly proportional to the original data value
[35]	Multiplicative	No	Evaluated linear and non-linear schemes
[36]	Additive$^+$	Yes	Showed that data hiding and additive noise could hinder consumer attributes idetification
[37]	Additive$^+$	Yes	Used uniform noise followed by homomorphic encryption
[38]	Additive$^+$	No	Either added random noise or augmented the data with fake noisy samples
[39]	Additive$^+$	No	Studied obfuscation and anonymization, considered privacy but not utility
[40]	Miscellaneous	Yes	Used a sparse dictionary
[41]	Miscellaneous	No	Used GANs for privacy-reservation of mobile datasets
[42]	Miscellaneous	Yes	Studied linear and data shuffling masking methods that required knowledge of the whole data set
[43]	Miscellaneous	Yes	Replaced high-risk data with alternative data

proportional to the size of the original value. In [35], the authors evaluated linear and non-linear log-based multiplicative noise schemes for protecting the confidentiality of micro data. The authors showed that the linear scheme was good if the data disseminator wanted to make minor changes to the original data, in exchange for data security. The non-linear scheme destroyed data utility for some items but maintained it in log-scale.

C. ADDITIVE$^+$

The authors in [36] showed that by hiding some parts of the weekly consumption data of individual consumers, and adding artificial noise could decrease the accuracy of attribute identification, such as employment status, household family size, etc. In [37], the authors proposed a two-step scheme. First, an SM added uniform noise, followed by homomorphic encryption before sending the data to an aggregator. The authors also compared different distributions in terms of the output of different values falling in different intervals between $-\infty$ to $+\infty$ and concluded that the uniform distribution was the best one as it affected all values equally. In [38], the authors proposed a method to preserve the privacy of training data in machine learning applications. The authors either added random noise to sensitive samples, or augmented the dataset with fake noisy samples to change the statistical properties of groups of samples.

In [39], the authors considered data such that each user's data had a normal distribution with a different mean for each user. The authors considered both obfuscation and anonymization, and showed that higher levels of anonymization and obfuscation would thwart the adversary's ability to correctly identify the consumers. The authors did not consider utility.

D. MISCELLANEOUS

The authors in [40] proposed to obfuscate appliance consumption signature by constructing a sparse dictionary of appliances' consumption data. The obfuscated data were then generated by randomly assigning different sparse coefficients with different probabilities. However, the authors underlined that the obfuscated data might lose their utility. In [25], an SM data obfuscation method based on the Gaussian mixture model (GMM) was proposed. Unlike existing data obfuscation approaches, this technique did not require a large number of consumers. However, it was shown in [27] that an eavesdropper could identify the original SM data using clustering techniques, thus posing privacy risks.

In [41], the authors used the GANs for generating the privacy-preserved data for mobile data sets and protected the recovery of consumers' trajectories from publicly available distributions. The authors in [42] studied linear and data shuffling masking models, which required information on the entire data set. The authors in [43] proposed a privacy-aware obfuscation method for web data. The authors further proposed an adversarial machine learning technique that

combined differential privacy-based noise addition with the probabilistic method.

In this paper, we investigate in detail the light-weight SM data obfuscation technique, which does not require any additional or special infrastructure, other than an SM. Unlike the work in [18] that only considered additive Gaussian noise-based data obfuscation approach, we consider several random distributions and compare both additive and multiplicative approaches. To the best of our knowledge, this is the first time that such an evaluation has been carried out for a light-weight SM data obfuscation method. Although the authors in [37] compared different distributions, our work is different from two aspects:

1) The work in [37] considered additive noise followed by homomorphic encryption-based obfuscation, whereas our work considers the performance of additive and multiplicative noise-based light-weight data obfuscation techniques.

2) The authors in [37] compared different distributions by only considering the values falling in different intervals. We carry out a detailed analysis by considering the change in distributions after data obfuscation as well as after calculation of mean, and provide an approximate value of variance and number of SM consumers required to achieve a reasonable level of privacy and utility performance, respectively.

In the following sections, we use a bold symbol such as $\boldsymbol{x}_p$ to denote a vector, whereas we use a non-bold symbol such as x_p^q to refer to the q^{th} single entry of $\boldsymbol{x}_p$. Other symbols such as X, x, X_n and x_n denote a scalar. Table 2 provides the list of symbols used in this paper.

III. SYSTEM MODEL

The SM scenario considered in this paper is shown in Fig. 1. We consider a group of M consumers, each of which has an SM installed at their premises. These consumers are supplied with energy by the ES, to which they also transmit their energy consumption several times in an hour. The accumulated energy consumption at the end of a billing period is used for billing purpose, while the periodic energy consumption received from each consumer is accumulated at the ES to understand the daily/hourly consumption patterns of the consumer group.

We consider that an SM $s_m, m = 1, 2, \ldots, M$ transmits its readings at evenly-spaced time-instants given by the set $t = t_1, t_2, \ldots, t_N$. The readings measured by the m^{th} SM at the n^{th} time-instant is given by r_n^m. The ES is interested in obtaining the sum of the readings at the n^{th} time-instant from all the consumers, i.e., $R_n = \sum_{m=1}^{M} r_n^m$, or the mean of the readings, i.e., $\mu_n = \frac{R_n}{M}$.

A consumer wants to hide his/her periodic electricity consumption from the ES. Similarly, the transmitted consumption should be protected from an eavesdropper, who can try to intercept these readings and gain insight into a user's behavior, or note his/her presence or absence in the premises. For this purpose, readings from each consumer are obfuscated such that the obfuscated readings $\tilde{r}_n^m$ are different from their original readings, i.e.,

$$\tilde{r}_n^m \neq r_n^m, \quad \forall m. \tag{1}$$

However, it should also be kept in mind that the ES should be able to accurately calculate the actual mean of the electricity consumption from these obfuscated readings, i.e.,

$$\tilde{\mu}_n \approx \mu_n, \tag{2}$$

where $\tilde{\mu}_n$ is the mean estimated from the obfuscated readings, i.e.,

$$\tilde{\mu}_n = \frac{\sum_{m=1}^{M} \tilde{r}_n^m}{M}. \tag{3}$$

Thus, (1) ensures privacy, i.e., the actual instantaneous consumption of the user is hidden from an eavesdropper, and (2) ensures utility, the actual instantaneous mean of the electricity consumed by M consumers can be calculated accurately from the obfuscated data. This data obfuscation can be carried out using random noise. The random nature of noise means that in the calculation of the mean from the obfuscated data, either the effect of noise would cancel out, or it would create a known constant offset in the calculated mean. However, to achieve an accurate estimate of the mean, the number of consumers over which the mean is calculated should be sufficiently large.

The data obfuscation approach should not require any extra infrastructure, so that it can be integrated in the current setup without any excessive additional expenditure or change of protocols. At the same time, it should be flexible so that any extra entity can be added in the future to provide an extra layer of security. Each SM should be able to carry out data obfuscation independently of other SMs without any intra-SM communication, and the communication from the ES to SMs should be kept to a minimum.

Our system model assumes that each SM records and reports its obfuscated readings truthfully to the ES without any tampering, and all the SM readings are synchronized in time. The obfuscation mechanism is fixed into the SM. It is considered that each SM has an accumulator that separately sums the total energy consumption at the end of the billing period and transmits it to the ES. We are only interested in obfuscating the periodic information sent by the SMs to the ES. The ES can know some of the parameters used for data obfuscation. Each SM has a unit for random number generation, as well as for carrying out multiplications and additions. An eavesdropper can note the transmitted readings, but not tamper with them. It is assumed that there is a single SM installed per consumer, although the proposed technique is not limited by this assumption.

In the following sections, we evaluate different noise distributions to be used in additive and multiplicative data obfuscation, as these techniques can be considered as lightweight approaches, which do not require the addition of any new infrastructure. We evaluate them in terms of the required parameters to achieve effective obfuscation, as well as the

TABLE 2. List of symbols.

Symbol	Description	Symbol	Description
$\boldsymbol{r}_n/\tilde{\boldsymbol{r}}_n$	Actual/Obfuscated data vector	$\gamma(.)$	Lower incomplete Gamma function
$\mu_n/\tilde{\mu}_n$	Mean of actual/obfuscated data vector	$\sigma_n/\tilde{\sigma}_n$	Standard deviation of actual/obfuscated data
f_i	i^{th} factor for choosing the obfuscation mechanism	$\mathcal{N}(.)/$ $\mathcal{N}_{GG}(.)$	Gaussian/generalized Gaussian distribution
M	Number of smart meters	$\chi^2(.)$	Chi-square distribution
k	Parameter of chi-square distribution	erf(.)	Error function
$\mu_{\xi_n}/\mu_{\zeta_n}$	Mean of additive/multiplicative noise vector	C_j	Cost criterion for the j^{th} obfuscation mechanism
$\sigma_{\xi_n}/\sigma_{\zeta_n}$	Standard deviation of additive/multiplicative noise vector	F	Number of considered factors for selecting an obfuscation mechanism
$\boldsymbol{\xi}_n/\boldsymbol{\zeta}_n$	Additive/Multiplicative noise vector	w_i	Relative weight for the i^{th} factor
$\Gamma(.)$	Gamma function	p_i	Penalty for the i^th factor
μ_{GG}, β_{GG}, ρ_{GG}	Mean, scale and shape parameters of generalized Gaussian noise	w	Width of the confidence interval of the distribution of the average obfuscated readings

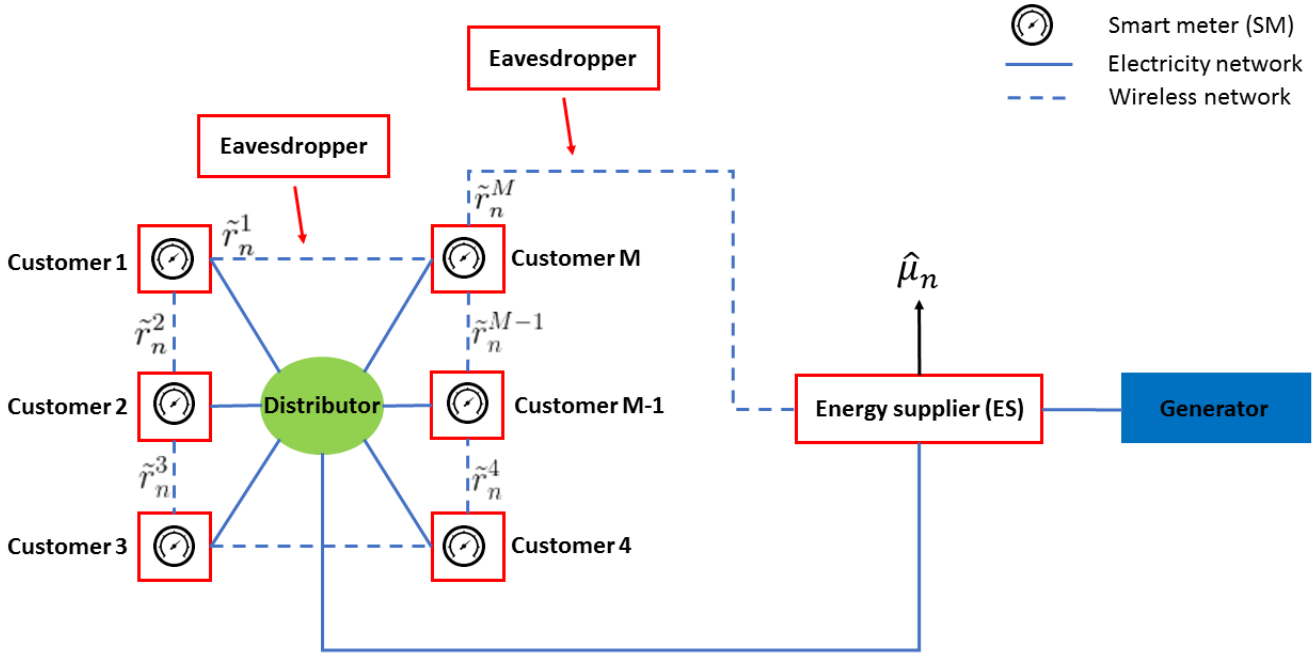

FIGURE 1. System model showing the obfuscated data transmitted to the ES by the consumers. The ES estimates the mean of the actual consumption from these data.

minimum number of consumers required to obtain accurate estimate of the mean value.

IV. ADDITIVE AND MULTIPLICATIVE NOISE BASED DATA OBFUSCATION AT SMART METERS

First, we consider a data obfuscation method that adds noise, preferably with zero-mean where possible, to the actual readings. According to this model, the new readings after the addition of noise are given as

$$\tilde{\boldsymbol{r}}_n = \boldsymbol{r}_n + \boldsymbol{\xi}_n, \tag{4}$$

where $\tilde{\boldsymbol{r}}_n$ represents the obfuscated readings at a time-instant t_n for all the smart meters, and $\boldsymbol{\xi}_n$ is the additive noise vector at that time. The mean and standard deviation of the noise are represented by μ_{ξ_n} and σ_{ξ_n}, respectively. The corresponding mean of the obfuscated data is

$$\tilde{\mu}_n = \mu_n + \mu_{\xi_n}, \tag{5}$$

and the standard deviation is given as

$$\tilde{\sigma}_n = \sigma_{\xi_n}. \tag{6}$$

The obfuscated readings are then sent to the ES that calculates the mean of $\tilde{\boldsymbol{r}}_n$.

We also consider multiplicative noise based SM data obfuscation, where the original data are obfuscated by multiplication with a random noise vector $\boldsymbol{\zeta}_n$, i.e.,

$$\tilde{\boldsymbol{r}}_n = \boldsymbol{r}_n \odot \boldsymbol{\zeta}_n, \tag{7}$$

where $\odot$ represents element-by-element multiplication. The mean and standard deviation of this noise vector are given by μ_{ζ_n} and σ_{ζ_n}, respectively. The mean of the obfuscated readings is given as

$$\tilde{\mu}_n = \mu_n \mu_{\zeta_n}, \tag{8}$$

and the standard deviation is given as [44]

$$\tilde{\sigma}_n = \sqrt{(\sigma_{\zeta_n}^2 + \mu_{\zeta_n}^2) \times (\sigma_n^2 + \mu_n^2) - (\mu_n \mu_{\zeta_n})^2}. \tag{9}$$

When zero-mean noise distribution is used for multiplicative data obfuscation, (6) simplifies to

$$\tilde{\sigma}_n = \mu_n \sigma_{\zeta_n}. \tag{10}$$

Based on the details presented in [18], it is clear that effective data obfuscation is achieved if there is a high probability that a reading received by an eavesdropper is different from the actual reading. This probability can be measured by the width of the 50% confidence interval of the obfuscated data distribution, where a large width would ensure that the obfuscated reading is different from the actual reading with a high probability. Given the mean of the readings transmitted by M SMs as equal to μ_n, we consider that noise should be added such that the width of the 50% confidence interval is at least $2\mu_n$, which means that either

- at least half of the obfuscated readings would fall outside the interval of $[0, 2\mu_n]$ or
- at least half of the obfuscated readings would fall outside the interval of $[-\mu_n, +\mu_n]$

The former criterion is used when the noise distribution only consists of positive samples, and the latter criterion is used when the noise distribution consists of both positive and negative values. The standard deviation of the required obfuscation noise can be calculated by either using the formula for the confidence interval, or the cumulative distribution function (CDF) of the obfuscated data's distribution.

Next, we calculate the required standard deviation for additive and multiplicative data obfuscation using Gaussian, Rayleigh, generalized Gaussian and chi-square distributions.

1) ADDITIVE GAUSSIAN DISTRIBUTION

As shown in [18], zero-mean random Gaussian noise, i.e., $\boldsymbol{\xi}_n \sim \mathcal{N}(0, \sigma_{\xi_n})$ can be added to obfuscate the readings. Equation (35) given in the Appendix can be used to calculate the required standard deviation, which is given as

$$\sigma_{\xi_n} = \frac{\mu_n}{0.6745}. \tag{11}$$

This approach is presented in [18] and used as the benchmark model in this paper.

2) ADDITIVE RAYLEIGH DISTRIBUTION

The Rayleigh distribution is made up of two components as follows:

$$\overline{\boldsymbol{\xi}}_n \sim \left| \mathcal{N}\left(0, \frac{\sigma_{\xi_n}}{\sqrt{2}}\right) + j\mathcal{N}\left(0, \frac{\sigma_{\xi_n}}{\sqrt{2}}\right) \right|, \tag{12}$$

where $\mathcal{N}(0, \frac{\sigma_{\xi_n}}{\sqrt{2}})$ is a distribution with mean equal to zero and standard deviation equal to $\frac{\sigma_{\xi_n}}{\sqrt{2}}$, i.e., each component making up the distribution has a standard deviation of $\frac{\sigma_{\xi_n}}{\sqrt{2}}$. The mean of the resulting distribution is [45]

$$\mu_{\overline{\xi}_n} = \frac{1.253}{\sqrt{2}} \sigma_{\xi_n}, \tag{13}$$

and the standard deviation is [45]

$$\sigma_{\overline{\xi}_n} = \frac{0.655}{\sqrt{2}} \sigma_{\xi_n}. \tag{14}$$

The required standard deviation can be calculated using (36), which results in

$$\sigma_{\xi_n} = \frac{2\mu_n}{\sqrt{-\ln(0.5)}}, \tag{15}$$

that can be used to generate the noise for SM data obfuscation.

3) ADDITIVE GENERALIZED GAUSSIAN DISTRIBUTION

Generalized Gaussian distribution depends on parameters mean μ_{GG}, scale β_{GG} and shape ρ_{GG}. It is written as [46]

$$\boldsymbol{\xi}_n \sim \mathcal{N}_{GG}(\mu_{GG}, \beta_{GG}, \rho_{GG}). \tag{16}$$

We consider a zero-mean generalized Gaussian noise, i.e., $\mu_{GG} = 0$, because a zero-mean additive noise does not require the ES to have any knowledge of the data obfuscation parameters for estimating the mean. The standard deviation of the distribution is [46]

$$\sigma_{\xi_n} = \frac{1}{\sqrt{\beta_{GG}}} \sqrt{\frac{\Gamma\left(\frac{3}{\rho_{GG}}\right)}{\Gamma\left(\frac{1}{\rho_{GG}}\right)}}, \tag{17}$$

where $\Gamma(.)$ is the Gamma function. The required standard deviation can be calculated using (37), which simplifies to

$$\frac{\gamma\left(\frac{1}{\rho_{GG}}, |\mu_n \sqrt{\beta_{GG}}|^{\rho_{GG}}\right)}{\Gamma\left(\frac{1}{\rho_{GG}}\right)} = 0.5. \tag{18}$$

In (18), $\gamma(.)$ is the lower incomplete Gamma function. Keeping ρ_{GG} fixed, we can solve (18) to get the value of β_{GG} required to achieve effective obfuscation.

4) ADDITIVE CHI-SQUARE DISTRIBUTION

Chi-square distribution depends on the parameter k and is expressed as follows:

$$\boldsymbol{\xi}_n \sim \chi^2(k/2). \tag{19}$$

The mean of the noise generated from the above distribution is [47]

$$\mu_{\xi_n} = k, \tag{20}$$

and the standard deviation of the distribution is [47]

$$\sigma_{\xi_n} = \sqrt{2k}. \tag{21}$$

The value of k required to achieve effective obfuscation can be obtained using the expression of the CDF of the Chi-square distribution in ((38)). The result is

$$\frac{\gamma\left(\frac{k}{2}, \mu_n\right)}{\Gamma\left(\frac{k}{2}\right)} = 0.5, \tag{22}$$

which can be used to calculate the required value of k.

5) MULTIPLICATIVE GAUSSIAN DISTRIBUTION

We consider a Gaussian distribution having zero mean and a standard deviation of σ_{ζ_n}, i.e., $\boldsymbol{\zeta}_n \sim \mathcal{N}(0, \sigma_{\zeta_n})$. Using (10) and (39), the required standard deviation value of the noise is equal to

$$\sigma_{\zeta_n} \approx \frac{1}{0.48\sqrt{2}}. \tag{23}$$

6) MULTIPLICATIVE RAYLEIGH DISTRIBUTION

Multiplicative data obfuscation using a non-zero mean noise vector will ensure that the mean of the original data is shifted. This shifting increases the probability of an eavesdropper receiving a reading different from the actual reading, and consequently, data obfuscation can be carried out by using a noise with a smaller standard deviation, compared to (35). In turn, the use of noise with a smaller standard deviation facilitates using a lower number of SMs. Motivated by this, we consider using noise with a Rayleigh distribution to carry out multiplicative data obfuscation. The distribution is given as follows:

$$\boldsymbol{\zeta}_n \sim \left|\mathcal{N}\left(0, \frac{\sigma_{\xi_n}}{\sqrt{2}}\right) + j\mathcal{N}\left(0, \frac{\sigma_{\xi_n}}{\sqrt{2}}\right)\right|. \tag{24}$$

The CDF of the Rayleigh distribution can be used to calculate the expression required to achieve effective data obfuscation, as shown in (40). Solving (40) gives

$$\sigma_{\xi_n} = \frac{2}{\sqrt{-\ln(0.5)}}, \tag{25}$$

which can be used for SM data obfuscation.

7) MULTIPLICATIVE GENERALIZED GAUSSIAN DISTRIBUTION

In this case, noise is generated according to (16). To calculate the required standard deviation, (18) can be modified to

$$\frac{\gamma\left(\frac{1}{\rho_{GG}}, |\sqrt{\beta_{GG}}|^{\rho_{GG}}\right)}{\Gamma\left(\frac{1}{\rho_{GG}}\right)} = 0.5.$$

Keeping ρ_{GG} fixed, we can solve to get the value of β_{GG} required to achieve effective obfuscation, which can then be used in (42) to calculate the required standard deviation.

8) MULTIPLICATIVE CHI-SQUARE DISTRIBUTION

Using the expression of the CDF of the distribution of the obfuscated signal in (46), we can obtain the following equation:

$$\frac{\gamma\left(\frac{k}{2}, 1\right)}{\Gamma\left(\frac{k}{2}\right)} = 0.5. \tag{26}$$

Subsequently, we can solve (26) to get the value of k required to achieve effective obfuscation.

V. CALCULATION OF MEAN AT THE ENERGY SUPPLIER

The ES generates an estimated mean of the actual readings. This estimate should be close to the actual average value μ_n with a high probability. This is ensured by having a high probability of the estimated mean to be close to the actual mean, which can be measured by a very small width w of the confidence interval of the distribution of the average obfuscated readings. We consider that the 99.5% confidence interval of the resulting distribution of the averaged result is equal to $w = 0.005\mu_n$. Next, we describe the calculation of mean and the minimum number of SMs required to obtain an accurate estimation for both additive and multiplicative noise.

A. ADDITIVE NOISE BASED DATA OBFUSCATION

The obfuscation readings are summed to calculate the mean. If the additive noise used has a zero-mean, the mean of the original data is equal to that of the obfuscated data; otherwise, the mean of the obfuscation noise should be subtracted from the estimated mean. Using the Central Limit Theorem, the calculated mean is considered to be normally distributed with mean μ_n and standard deviation $\frac{\tilde{\sigma}_n}{\sqrt{M}}$. Consequently, using (47), the number of meters to estimate the mean accurately should be

$$M = \left(\frac{2.81\tilde{\sigma}_n}{0.005\mu_n}\right)^2, \tag{27}$$

for Gaussian, generalized Gaussian and chi-square distributions. In the case of Rayleigh distribution, the number of meters can be written as

$$M = \left(\frac{2.81\sigma_{\bar{\xi}_n}}{0.005\mu_n}\right)^2. \tag{28}$$

B. MULTIPLICATIVE NOISE BASED DATA OBFUSCATION

In case of multiplicative noise, the mean of the actual readings is calculated in two different ways depending on the types of noise distributions used:

1) MULTIPLICATIVE GAUSSIAN AND GENERALIZED GAUSSIAN

As the mean of the noise is zero, the estimated mean is calculated by dividing the standard deviation of the obfuscated readings by that of the obfuscation noise, i.e., $\hat{\mu}_n = \frac{\tilde{\sigma}_n}{\sigma_{\zeta_n}}$. Using the Central Limit Theorem, the calculated mean will be normally distributed with mean μ_n. On the other hand, the standard deviation of the calculated mean can be considered as standard deviation of the samples of standard deviation, known as the standard error of standard deviation [48]. It is equal to $\frac{\tilde{\sigma}_n}{\sqrt{2M}\sigma_{\zeta_n}}$ for multiplicative Gaussian noise.

As described in the previous section, a value of $w = 0.005\mu_n$ for the confidence interval of the standard error ensures that an accurate estimate of the mean is obtained from the obfuscated readings. Using this value and the expression

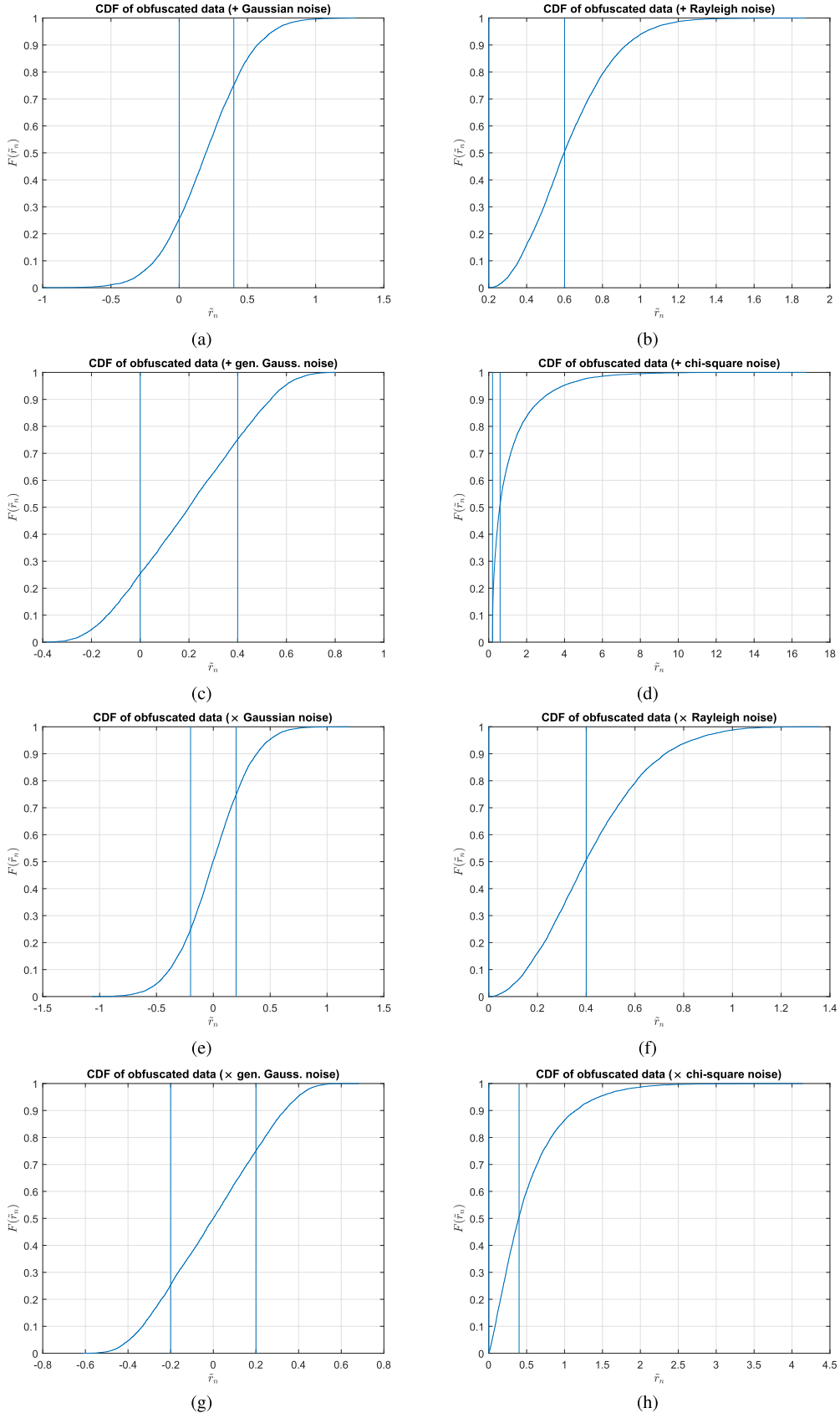

FIGURE 2. Cumulative distribution obtained from simulated obfuscated data (a) additive Gaussian (b) additive Rayleigh (c) additive generalized Gaussian (d) additive chi-square (e) multiplicative Gaussian (f) multiplicative Rayleigh (g) multiplicative generalized Gaussian (h) multiplicative chi-square.

of the standard error in (47), the number of SMs required to achieve the necessary estimation accuracy is

$$M = \left(\frac{2.81\tilde{\sigma}_n}{0.005\sqrt{2}\mu_n\sigma_{\zeta_n}} \right)^2, \tag{29}$$

which can be simplified to

$$M = \left(\frac{2.81}{0.005\sqrt{2}} \right)^2. \tag{30}$$

Using the Central Limit Theorem, the calculated mean in the case of generalized Gaussian is considered to be normally distributed with mean μ_n and standard deviation $\frac{\tilde{\sigma}_n}{\sqrt{3.7M}\sigma_{\zeta_n}}$, where the value of 3.7 is derived empirically, as described in the Appendix. Consequently, using (47), the number of meters to estimate the mean accurately should be

$$M = \left(\frac{2.81\tilde{\sigma}_n}{0.005\sqrt{3.7}\sigma_{\zeta_n}\mu_n} \right)^2, \tag{31}$$

which can be simplified as

$$M = \left(\frac{2.81}{0.005\sqrt{3.7}} \right)^2. \tag{32}$$

2) MULTIPLICATIVE RAYLEIGH AND CHI-SQUARE

In this case, the mean of noise is non-zero, and the estimated mean is calculated by dividing the mean of the obfuscated readings by that of the obfuscation noise, i.e., $\hat{\mu}_n = \frac{\tilde{\mu}_n}{\mu_{\zeta_n}}$. Using the Central Limit Theorem, the calculated mean for both multiplicative Rayleigh and Chi-Square data obfuscation is considered to be normally distributed with mean μ_n and standard deviation $\frac{\tilde{\sigma}_n}{\sqrt{M}\mu_{\zeta_n}}$. Consequently, the number of meters to estimate the mean accurately can be calculated using (47) as follows:

$$M = \left(\frac{2.81\tilde{\sigma}_n}{0.005\mu_n\mu_{\zeta_n}} \right)^2. \tag{33}$$

VI. VALIDATION AND ANALYSIS

We consider an average reading μ_n equal to 0.2 kW and calculate the parameters required to achieve effective obfuscation, as well the minimum number of consumers required to achieve an accurate estimate of the mean value. We carry out these calculations for each obfuscation approach presented in Sections III and IV. For additive Gaussian obfuscation, which is the approach presented in [18], σ_{ξ_n} is equal to 0.3 kW. To accurately calculate the mean of the consumers' readings, M should be high, e.g., around 700000 consumers. This approach is used as a reference for comparing to other approaches, and is called the benchmark approach in the following discussion.

A. PARAMETERS AND COMPARISON OF M WITH THE BENCHMARK APPROACH

1) ADDITIVE APPROACHES

According to (15), the standard deviation of the Rayleigh noise σ_{ζ_n} used in (12) and the standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings equal approximately to 0.2225 kW. Using this value in (28), the minimum number of required SMs is approximately equal to 390000, which is about 40% less than that required by the benchmark approach.

Considering $\rho_{GG} = 5$, according to (37), $\beta_{GG} = 5.31$, the standard deviation σ_{ξ_n} of the generalized Gaussian noise in (17) and the standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings is approximately equal to 0.2472 kW. Using this value in (27), the number of required SMs is approximately equal to 480000, which is about 30% less than that required by the benchmark approach.

According to (38), $k = 0.93$, and the standard deviation σ_{ξ_n} of the chi-square distributed noise in (21) and the standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings is approximately equal to 1.3638 kW. Using this value in (27), the minimum number of required SMs is approximately equal to 15 million, which is phenomenally higher than that required by the benchmark approach. This high number is due to the large standard deviation of the obfuscated SM readings.

2) MULTIPLICATIVE APPROACHES

According to (23), the standard deviation σ_{ζ_n} of the Gaussian noise used in (7) is approximately equal to 1.473 kW. The standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings equals 0.295 kW. Using this value in (29), the number of required SMs is approximately equal to 158000, which is about 80% less than that required by the benchmark approach.

According to (14) and (25), the standard deviation σ_{ζ_n} of the Rayleigh noise used in (7) is approximately equal to 1.1 kW. However, the standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings is equal to 0.22 kW. Using this value in (33), the number of SMs required comes out to be about 85000, which is approximately 90% less than that required by the benchmark approach.

Considering $\rho_{GG} = 5$, according to (43), $\beta_{GG} = 0.2122$, the standard deviation σ_{ζ_n} of the generalized Gaussian noise in (42) and the standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings are approximately equal to 1.2364 kW and 0.2472 kW, respectively. Using these values in (31), the number of required SMs is approximately equal to 85000, which is about 90% less than that required by the approach presented in [18].

According to (46), $k = 2.6$, and the standard deviation σ_{ζ_n} of the chi-square multiplicative noise and the standard deviation $\tilde{\sigma}_n$ of the obfuscated SM readings are approximately equal to 2.2804 kW and 0.456 kW, respectively. Using these values in (33), the number of required SMs is approximately equal to 243000, which is about 65% less than that required by the benchmark approach.

B. OBFUSCATED DATA AND CALCULATED MEAN

Next, we use the parameters calculated in the previous sub-sections to simulate the SM scenario shown in Fig. 1. We generate noise according to these parameters, and obfuscate the SM data with the generated noise. We carry out this process for a total of 1000 times, and plot the cumulative

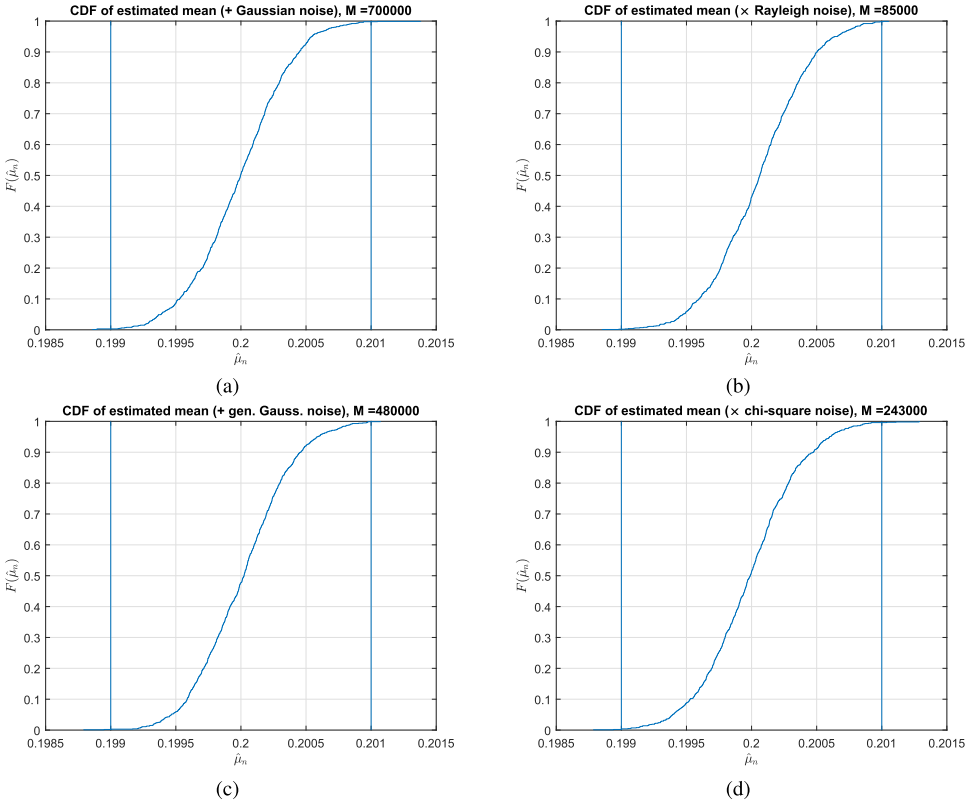

FIGURE 3. Cumulative distribution obtained from calculated mean (a) additive Gaussian (b) multiplicative Rayleigh (c) additive generalized Gaussian (d) multiplicative chi-square.

distribution of the resulting obfuscated data. We used Matlab R2019a to perform the simulations on a system with Intel i5-4200U processor and 12 GB RAM.

The results are shown in Fig. 2, where Figs. 2(a) - 2(d) represent additive noise data obfuscation, and Figs. 2(e) - 2(h) represent multiplicative noise data obfuscation. We also plot two parallel lines placed 0.4 kW apart to show the corresponding cumulative probability at these values. The difference of probability values in between these lines should be 0.5, which shows that 50% of the obfuscated data lie within an interval of 0.4 kW. All the simulated results show that 50% of the obfuscated data lie within these lines, thus confirming that the calculated parameters are correct. Note that the data obfuscated using Gaussian and generalized Gaussian noise consist of both positive and negative values, and the data obfuscated using the chi-square distribution is spread over large values, especially for the additive case.

Next, we calculate the mean of the obfuscated data over the number of SMs calculated in the previous sub-sections for each approach. We again carry out the simulations over 1000 iterations and plot the corresponding cumulative distribution of the calculated mean values. As all the results are identical, we only show four figures in Fig. 3. Two parallel lines are placed corresponding to 0.201 kW and 0.199 kW, which represent an interval such that the minimum and maximum boundaries of the interval are lower and higher than 0.5% of 0.2 kW. The difference between the probability values shown on the y-axis at intersection of these two boundaries and the cumulative distribution plot should be 0.995, i.e., very close to 1. It can be clearly observed that most of the readings are very close to 0.2 kW and fall inside the area of the parallel lines, underlining the fact that the number of SMs is sufficient to accurately estimate the mean. The multiplicative data obfuscation using Rayleigh distribution requires the least number of SMs.

As the minimum number of SMs is 85000, required by the Rayleigh and generalized Gaussian distributions based multiplicative data obfuscation, we also show the mean estimated using 85000 SMs for each data obfuscation approach. The results are plotted in Fig. 4. As it can be observed, the area between the two parallel lines decreases, meaning that the accuracy of the estimated mean decreases. The additive approaches are affected the most, because they require a higher number of SMs, compared to the multiplicative approaches. The additive chi-square approach is affected the most, followed by additive Gaussian noise based approach. The additive Rayleigh and generalized Gaussian approaches are affected almost equally. As noted earlier, the multiplicative approaches are not significantly affected. The multiplicative chi-square noise is affected the most, followed by the multiplicative Gaussian approach.

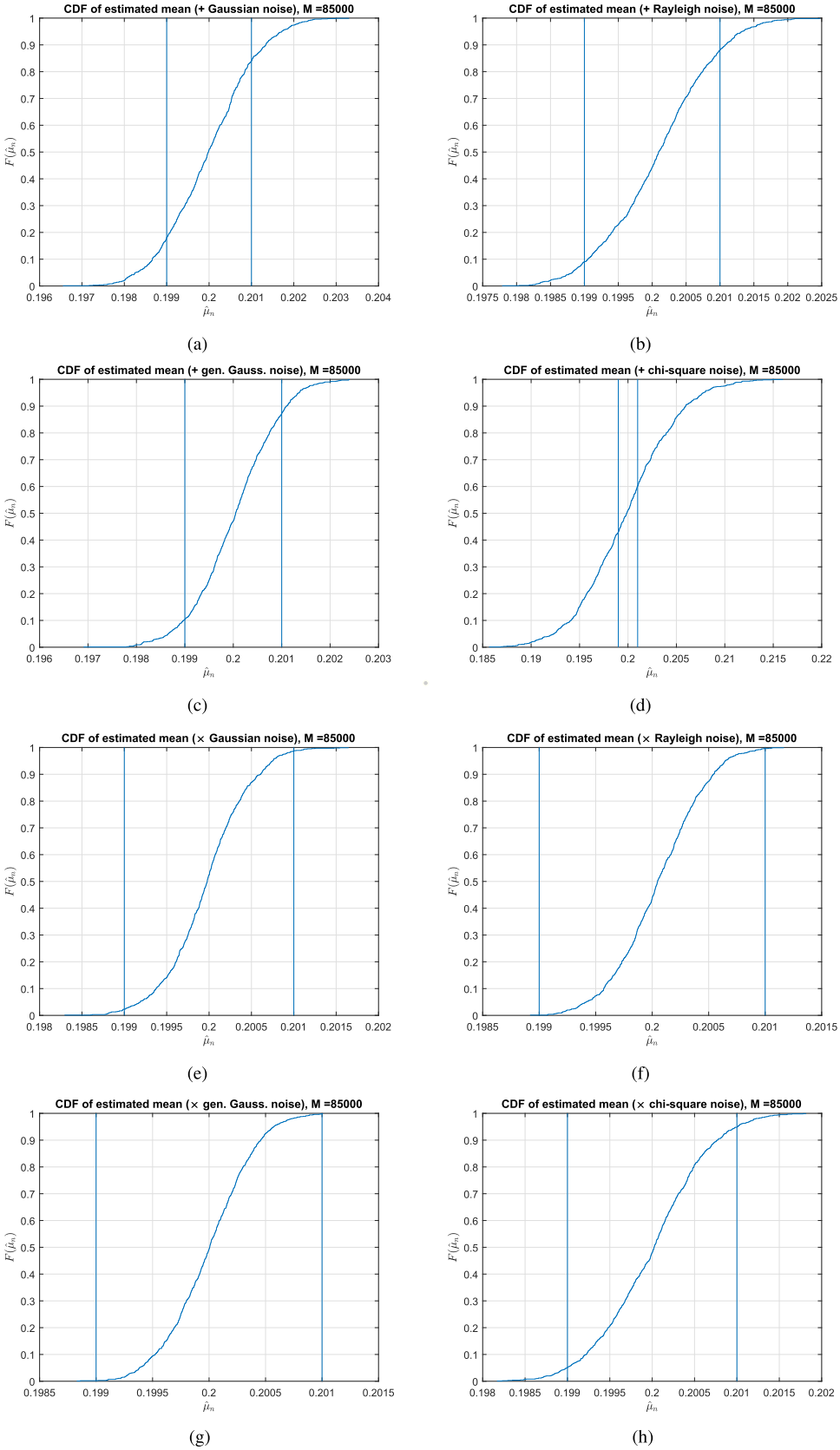

FIGURE 4. Cumulative distribution obtained from calculated mean for M = 85000 (a) additive Gaussian (b) additive Rayleigh (c) additive generalized Gaussian (d) additive chi-square (e) multiplicative Gaussian (f) multiplicative Rayleigh (g) multiplicative generalized Gaussian (h) multiplicative chi-square.

As mentioned in [49], Laplace and Cauchy noise are commonly found in practical applications. The former noise distribution has also been used for additive SM data obfuscation, as shown in Table 1. Therefore, we find it pertinent to comment on the use of these distributions for additive and multiplicative SM data obfuscation. Our simulations show that the minimum numbers of SMs required to accurately recover the mean of the SM readings should be about 500000 and 200000 for additive and multiplicative Laplace distribution based SM data obfuscation, respectively.

The required number of SMs for additive Laplace noise is the second highest after that for chi-square distribution, and almost the same as that for generalized Gaussian distribution. The number of SMs for multiplicative Laplace noise is the second highest after that required for chi-square distribution. This behavior shows that the performance of the Laplace distribution, if used for SM data obfuscation, will be between that of chi-square and generalized Gaussian distributions. On the other hand, the mean and standard deviation of Cauchy distribution do not exist [50]. As the main constraint on SM data obfuscation is to be able to accurately recover the mean of the SM readings for a group of consumers, it is not possible to use this distribution for SM data obfuscation.

C. DATA OBFUSCATION WITH VARYING MEAN

As the mean of the readings may vary over time, and the obfuscation noise is calculated for a fixed mean, we show the impact of fixed noise on data obfuscation. The data are obfuscated according to an assumed value of 0.2 kW, while the mean is expected to vary by ± 0.04 kW, i.e., a $\pm 20\%$ variation. We first show the results for $+0.04$ kW variation in Fig. 5. An impact is expected to increase the area in between the two parallel lines, because the obfuscation noise is calculated for a mean that is lower than the actual mean. This will not impact the estimation of the mean, but would rather affect the privacy level, meaning that there is a higher chance of the privacy of a customer being invaded. The additive approaches are affected, whereas the multiplicative approaches are not affected, because the latter are independent of the value of the readings being obfuscated, and depend only on the desired obfuscation level. This can also be confirmed by the expressions of the obfuscation noise's variance calculated in Sections III and IV. It appears that the additive chi-square noise is affected less than the other approaches.

We show similar results where the actual mean is 0.16 kW, while the noise parameters are calculated for a mean of 0.2 kW. The results are shown in Fig. 6. As expected, the area between the two parallel lines for the additive approaches decreases as the actual obfuscation is less than the desired level. A higher number of obfuscated data samples will lie outside the 0.2 kW ± 0.2 kW interval and consequently, this higher percentage of obfuscated readings would degrade the estimated mean's accuracy, or would require a higher number of SMs to estimate the mean. The additive chi-square approach is affected the least, followed by the additive Rayleigh data obfuscation approach. As expected, the multiplicative approaches are not affected by a change in the mean value.

These results show that for the additive approaches, the SMs should generate noise according to the changing mean. Thus, the ES should send the value of an estimated mean to the SMs at periodic intervals or whenever it estimates the current obfuscation mechanism to be insufficient. Another option is to always use a high standard deviation of obfuscated than that was calculated in Section III, which will require the use of a higher number of SMs.

D. OBSERVATIONS

We can make the following observations based on the results:

- Multiplicative approaches in general require a lower number of SMs to accurately calculate the mean.
- Multiplicative approaches can carry out same level of data obfuscation, irrespective of the actual mean. Additive approaches, on the other hand, either require an estimate of the expected mean to carry out effective data obfuscation, or require more than the calculated number of SMs to avoid any decrease in the estimation accuracy.
- Multiplicative Rayleigh and generalized Gaussian data obfuscation approach require the least number of SMs, meaning that they are suitable for smaller neighborhoods compared to other approaches.
- The zero-mean additive approaches, do not require any information by the ES to calculate the mean. The non-zero mean additive approaches require the knowledge of the mean value.
- The zero-mean multiplicative data obfuscation approaches require the ES to be aware that the mean can be estimated by the standard deviation of the received obfuscated readings. The ES should also have knowledge about the obfuscation noise's standard deviation. The non-zero multiplicative approaches, on the other hand, require the value of the mean of the obfuscation noise used to obfuscate the data. Thus, these approaches require the ES to have more knowledge of the obfuscation mechanism, compared to the additive approaches.

Smart meter data obfuscation is practically applicable in current scenarios to protect the privacy of consumers. It does not require any particular new infrastructure, and only requires SMs equipped with a small processing unit. Current SMs are equipped with this feature, while the conventional meters can also be replaced by this type of SMs once they are phased out. At the ES, there is no particular hardware requirement for the calculation of mean from the obfuscated data provided by a group of consumers. The calculated mean can be used by grid operators, distributors and suppliers to provide different services, as outlined in Section I. Furthermore, the SMs, once installed, can be integrated along with new privacy-preserving methods by providing an extra layer of privacy.

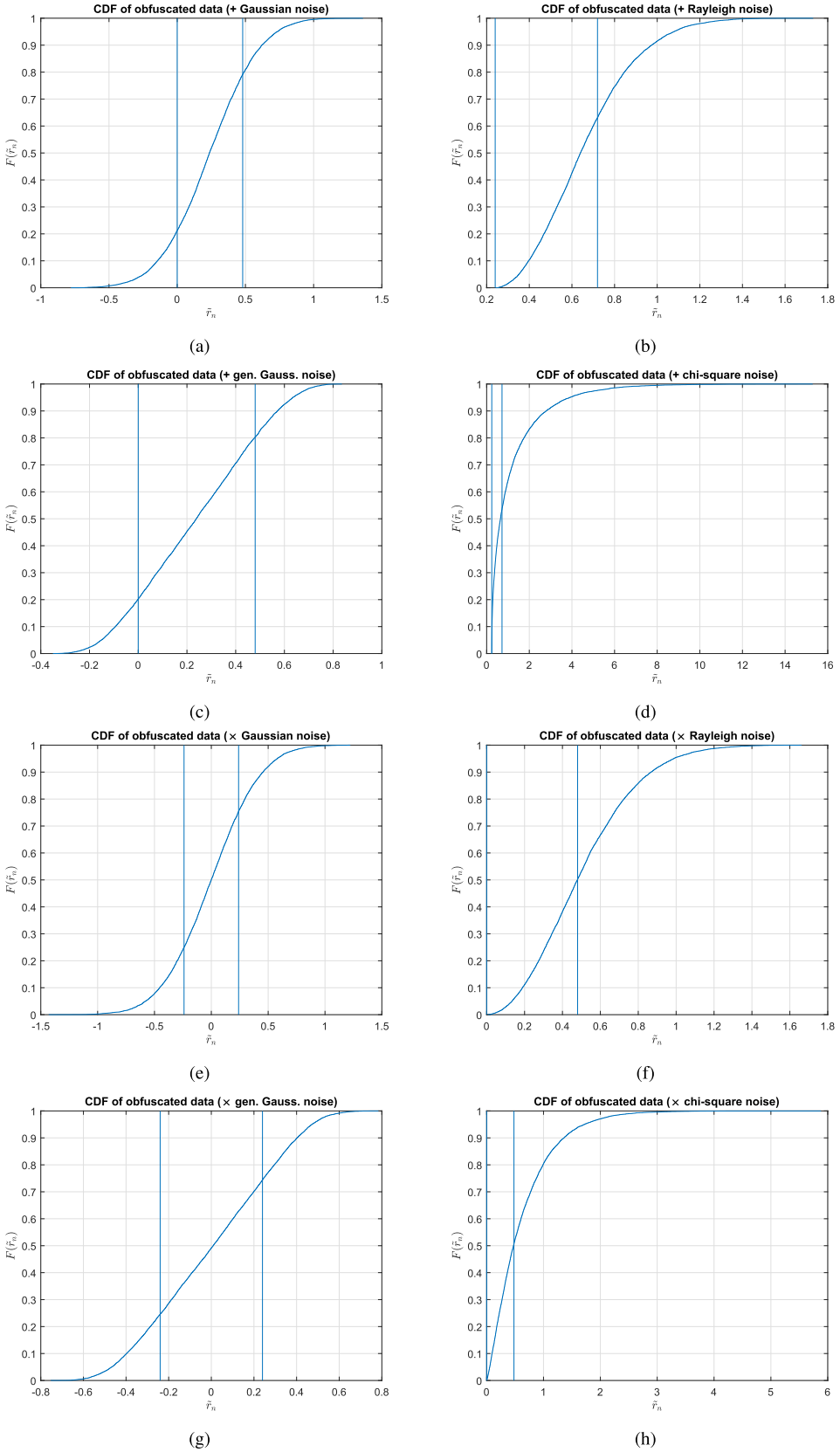

FIGURE 5. Cumulative distribution obtained from simulated obfuscated data with mean = 0.24 (a) additive Gaussian (b) additive Rayleigh (c) additive generalized Gaussian (d) additive chi-square (e) multiplicative Gaussian (f) multiplicative Rayleigh (g) multiplicative generalized Gaussian (h) multiplicative chi-square.

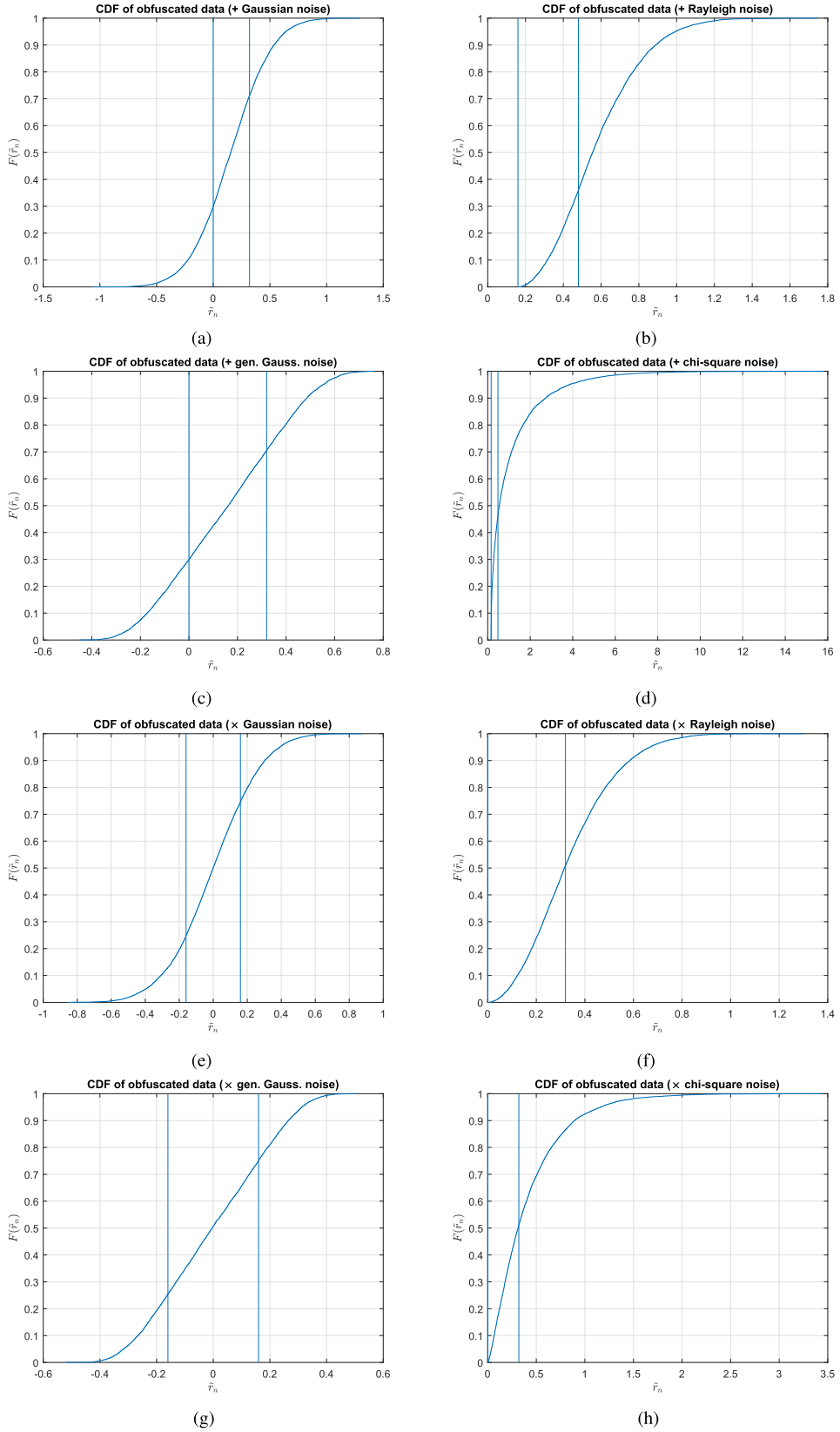

FIGURE 6. Cumulative distribution obtained from simulated obfuscated data with mean = 0.16 (a) additive Gaussian (b) additive Rayleigh (c) additive generalized Gaussian (d) additive chi-square (e) multiplicative Gaussian (f) multiplicative Rayleigh (g) multiplicative generalized Gaussian (h) multiplicative chi-square.

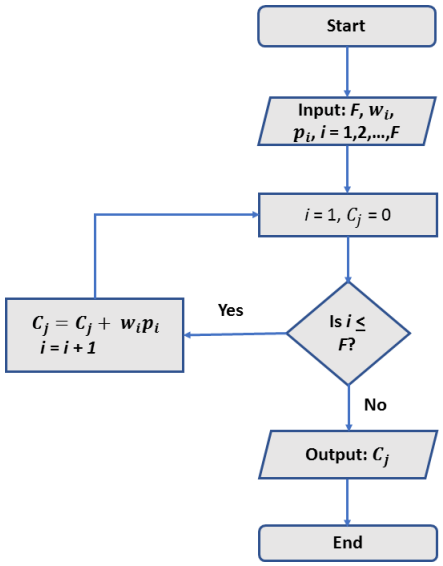

FIGURE 7. Procedure for calculating the cost for the j^{th} obfuscation mechanism.

E. GUIDELINES FOR CHOOSING AN OBFUSCATION MECHANISM

A cost criterion is used to select the most suitable data obfuscation mechanism. It is calculated based on $F = 5$ factors, represented as f_i, $i = 1, 2, \ldots, 5$. The individual penalty due to each factor is represented by p_i, $i = 1, 2, \ldots 5$. The cost criterion for the j^{th} obfuscation mechanism can be written as

$$C_j = \sum_{i=1}^{5} w_i p_i \tag{34}$$

where w_i is the relative weight given to each factor. The cost can be calculated for each obfuscation mechanism, and a lower cost is considered as better. Figure 7 illustrates the procedure of calculating the cost for the i^{th} obfuscation mechanism. Each individual penalty is assigned a value from 1, 2, 3 based on the requirement of each factor, where 1, 2 and 3 represent low, medium and high, respectively. Note that this assignment of value is relative. Based on Sections IV and V, we select the following factors for choosing a particular data obfuscation mechanism:

1) Computational complexity at the SM (f_1): We assume that the complexities of generating a random number, addition and multiplication are low and fixed. Thus, the computational complexity of data obfuscation at the SM does not vary with the choice of a particular noise distribution and obfuscation mechanism. The generation of required standard deviation needed to generate a random number requires division by constant numbers for Gaussian and Rayleigh distributions. However, the generalized Gaussian and chi-square distributions require calculation of β_{GG} and k, as shown in (18) and (22), respectively. Thus, the complexities are considered as high and medium for the generalized Gaussian and chi-square distributions, respectively, while the complexities of the remaining distributions are considered as low.
2) Computational complexity at the ES (f_2): The computational complexity for calculating the mean or standard deviation increases linearly as M increases. Thus, it is dependent on the minimum number of required SMs, and increases as this number increases. We consider the complexities as high for additive Gaussian, additive chi-square, medium for additive Rayleigh, additive generalized Gaussian, multiplicative Gaussian and multiplicative chi-square, and low for the remaining data obfuscation approaches.
3) Knowledge at the SM (f_3): The additive approaches require the SM to know an estimated value of the mean of the original data readings for calculating the standard deviation of the obfuscation noise. The requirement is classed as medium for these approaches. Additive generalized Gaussian distribution further requires the SM to know the value of ρ_{GG}. This information may need to be transmitted to each SM at a low frequency and the requirement is classed as high. The multiplicative approaches do not require any knowledge of the mean of the original data readings at the SM and the requirement of this factor is classed as low.
4) Knowledge at the ES (f_4): Zero-mean additive noise does not require any knowledge at the ES for calculation of mean and the requirement is considered as low. Non-zero additive and multiplicative obfuscation require knowledge of the mean, and knowledge of either the mean or standard deviation of noise and whether to calculate the mean or the standard deviation. Thus, for non-zero additive obfuscation, we consider the requirement as medium, while it is considered as high for multiplicative obfuscation.
5) Spread of values (f_5): Chi-square distribution is spread over large values, especially for multiplicative obfuscation. This means that high amplitude obfuscated readings will be generated, which requires a more complex modulation mechanism for transmission of the obfuscated data to the ES. The complexity is considered as medium and large for additive and multiplicative chi-square obfuscation, respectively. Other distributions are almost identically spread over a low range and are assigned a low complexity.

The individual penalties for each data obfuscation mechanism is represented by Table 3. In the table, + and × refer to additive and multiplicative obfuscation mechanisms, respectively, and G, R, GG and CS refer to Gaussian, Rayleigh, generalized Gaussian and chi-square distributions, respectively. Based on Table 3 and with all weights equal to 0.2, the costs are equal to 1.6, 1.8, 2, 2.4, 1.6, 1.4, 2.2, 2.4 for +G, +R, +GG, +CS, ×G, ×R, ×GG, ×CS, respectively. This mean that if all factors are equally important, multiplicative Rayleigh is the best obfuscation mechanism. As another example, if low computational complexity at the SM is very important, and that at the ES does not matter, we assign the weights 0.4 and

TABLE 3. Cost factor calculation.

Factor/Obfuscation	+G	+R	+GG	+CS	×G	×R	× GG	× CS
Computational complexity (SM)	1	1	3	2	1	1	3	2
Computational complexity (ES)	3	2	2	3	2	1	1	2
Required knowledge (SM)	2	2	3	3	1	1	3	2
Required knowledge (ES)	1	2	1	2	3	3	3	3
Spread of values	1	1	1	2	1	1	1	3

0 to f_1 and f_2, respectively. In this case, the costs are 1.2, 1.4, 2.2, 1.8, 1.4, 1.4, 2.6, 2.4, meaning that additive Gaussian is the best data obfuscation mechanism.

VII. CONCLUSION

In this paper, we presented and analyzed light-weight data obfuscation methods. These methods involved either additive or multiplicative noise. We evaluated the use of different noise distributions, including Gaussian, Rayleigh, generalized Gaussian and chi-square distributions. We first calculated the required noise parameters for achieving a given level of data obfuscation, as well as the minimum number of smart meters (SMs) required to accurately estimate the mean electricity consumption from the obfuscated data. We verified the calculations via simulations. Our results indicated that multiplicative noise based data obfuscation required a lower number of SMs compared to the additive noise based data obfuscation, and it was independent of the mean value of the actual consumption data. We further showed that SM data obfuscation using Rayleigh and generalized Gaussian distributions required the least number of SMs. We further presented guidelines for selecting the obfuscation mechanism according to various requirements. This work can be used in practical scenarios to calculate noise parameters according to a given data obfuscation level, and also define the size of the neighborhood required to estimate the mean at a given data obfuscation level.

APPENDIX

A. CALCULATION OF STANDARD DEVIATION

1) ADDITIVE GAUSSIAN DISTRIBUTION

To calculate the standard deviation of the noise required to achieve this level of data obfuscation, we follow the steps presented in [18]. We use the formula of the 50% confidence interval of a Gaussian distribution, which can be used to calculate standard deviation of the noise required to achieve effective data obfuscation:

$$\mu_n = 0.6745\sigma_{\xi_n}. \tag{35}$$

2) ADDITIVE RAYLEIGH DISTRIBUTION

To calculate the standard deviation of the required noise, we again consider that at least half of the readings should fall outside the interval $[0, 2\mu_n]$. As noise generated by (12) is always positive valued, it means that the probability of obfuscated readings being greater than $2\mu_n$ should be at least 50%. Making use of the formula for the CDF of the Rayleigh distribution, i.e., $F_R(\mu_n) = \exp\left(-\frac{0.5\mu_n^2}{(\sigma_{\xi_n}/\sqrt{2})^2}\right)$, we can calculate the required standard deviation by solving the following equation:

$$1 - F_R(2\mu_n) = 0.5. \tag{36}$$

3) ADDITIVE GENERALIZED GAUSSIAN DISTRIBUTION

To calculate the parameters of the required noise for achieving affective data obfuscation, we use the expression of the CDF of the generalized Gaussian distribution, i.e., $F_{GG}(\mu_n) = \frac{1}{2} + \frac{\text{sgn}(\mu_n)}{2\Gamma(1/\rho_{GG})}\gamma\left(\frac{1}{\rho_{GG}}, |\mu_n\sqrt{\beta_{GG}}|^{\rho_{GG}}\right)$, where sgn(.) is a function that gives the sign of its input and $\gamma(.)$ is the lower incomplete Gamma function [46]. At least half of the obfuscated readings should be greater than $|\mu_n|$. This is expressed as

$$F_{GG}(\mu_n) - F_{GG}(-\mu_n) = 0.5. \tag{37}$$

4) ADDITIVE CHI-SQUARE NOISE

To calculate the required value of k, we use the CDF of the chi-square distribution. The expression of the CDF is given as $F_{\chi^2}(\mu_n) = \frac{\gamma\left(\frac{k}{2}, \frac{\mu_n}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$ [47]. The parameter k should be such that at least half of the obfuscated readings are greater than $2\mu_n$, i.e.,

$$1 - F_{\chi^2}(2\mu_n) = 0.5. \tag{38}$$

5) MULTIPLICATIVE GAUSSIAN NOISE

To achieve effective obfuscation, the multiplicative noise should be such that at least 50% of the obfuscated readings should be greater than $|\mu_n|$. To calculate the required standard deviation of the noise for achieving this level of data obfuscation, we use the formula of the CDF of a normal distribution, which is denoted as $F_G(\mu_n)$. We get the following expression:

$$\begin{aligned} F_G(\mu_n) - F(-\mu_n) &= 0.5, \\ \text{erf}\left(\frac{\mu_n}{\tilde{\sigma}_n\sqrt{2}}\right) &= 0.5, \end{aligned} \tag{39}$$

where erf(.) is the error function.

6) MULTIPLICATIVE RAYLEIGH NOISE

The standard deviation of the noise components that will give readings greater than $2\mu_n$ with a 50% confidence interval can

be calculated by making use of the expression for the CDF of the Rayleigh distribution, i.e.,

$$1 - \exp\left(-0.5\frac{(2\mu_n)^2}{(\mu_n\sigma_{\xi_n}/\sqrt{2})^2}\right) = 0.5, \tag{40}$$

where the denominator in (40) is obtained because using (24) in (7), we get

$$\tilde{\boldsymbol{r}}_n \sim \left|\mathcal{N}\left(0, \frac{\mu_n\sigma_{\xi_n}}{\sqrt{2}}\right) + j\mathcal{N}\left(0, \frac{\mu_n\sigma_{\xi_n}}{\sqrt{2}}\right)\right|. \tag{41}$$

7) MULTIPLICATIVE GENERALIZED GAUSSIAN NOISE

The standard deviation of the obfuscated readings is as follows:

$$\begin{aligned}\tilde{\sigma}_n &= \mu_n\sigma_{\zeta_n}, \\ &= \frac{\mu_n}{\sqrt{\beta_{GG}}}\sqrt{\frac{\Gamma\left(\frac{3}{\rho_{GG}}\right)}{\Gamma\left(\frac{1}{\rho_{GG}}\right)}}.\end{aligned} \tag{42}$$

Subsequently, (37) is modified as

$$\frac{\gamma\left(\frac{1}{\rho_{GG}}, |\sqrt{\beta_{GG}}|^{\rho_{GG}}\right)}{\Gamma\left(\frac{1}{\rho_{GG}}\right)} = 0.5, \tag{43}$$

where the difference in the above expression compared to (37) is due to the presence of an additional term μ_n in the standard deviation of the obfuscated readings, compared to the standard deviation of obfuscated data in the presence of additive noise.

8) MULTIPLICATIVE CHI-SQUARE NOISE

To calculate the CDF of the obfuscated readings, we note that the CDF of the chi-square distribution can be written as follows in terms of a scale θ:

$$F_{\chi^2}(\mu_n) = \frac{\gamma\left(\frac{k}{2}, \frac{\mu_n}{\theta}\right)}{\Gamma\left(\frac{k}{2}\right)}, \quad \theta = 2. \tag{44}$$

Given the fact that now the mean and standard deviation of the obfuscated readings are equal to $k\mu_n$ and $\sqrt{2k}\mu_n$, the CDF of the obfuscated readings can be written as:

$$F^o_{\chi^2}(\mu_n) = \frac{\gamma\left(\frac{k}{2}, \frac{1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}, \quad \text{i.e., } \theta = 2\mu_n. \tag{45}$$

Thus, to find the parameter k such that 50% of the obfuscated readings fall outside the interval $[0, 2\mu_n]$, we get the following expression:

$$1 - F^o_{\chi^2}(2\mu_n) = 0.5. \tag{46}$$

B. CALCULATION OF MEAN

1) WIDTH OF THE CONFIDENCE INTERVAL

The width w of the 99.5% confidence interval required to accurately calculate the mean should be equal to [47]

$$w = \frac{2.81\sigma_{\xi_n}}{\sqrt{M}}. \tag{47}$$

2) STANDARD DEVIATION FOR MULTIPLICATIVE GENERALIZED GAUSSIAN DISTRIBUTION

As mentioned in Section V-B1, the calculated mean is considered to be normally distributed with mean μ_n and standard deviation $\frac{\tilde{\sigma}_n}{\sqrt{3.7M}\sigma_{\zeta_n}}$. The factor of 3.7 in the denominator is derived empirically by 1) generating random numbers following generalized Gaussian distribution for varying values of M, 2) estimating the mean of the generated random numbers for each value of M, 3) calculating the standard deviation of the estimated mean for each value of M, and 4) finding a factor f that matches the standard deviation values calculated using the random numbers with those calculated using the expression $\frac{\tilde{\sigma}_n}{\sqrt{fM}\sigma_{\zeta_n}}$.

REFERENCES

[1] M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac, "Smart meter data privacy: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2820–2835, Jun. 2017.

[2] S. Desai, R. Alhadad, N. Chilamkurti, and A. Mahmood, "A survey of privacy preserving schemes in IoE enabled smart grid advanced metering infrastructure," *Cluster Comput.*, vol. 22, no. 1, pp. 43–69, Mar. 2019.

[3] Z. Erkin, J. R. Troncoso-Pastoriza, R. L. Lagendijk, and F. Perez-Gonzalez, "Privacy-preserving data aggregation in smart metering systems: An overview," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 75–86, Mar. 2013.

[4] S. Finster and I. Baumgart, "Privacy-aware smart metering: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1088–1101, 2nd Quart., 2015.

[5] L. Zhu, Z. Zhang, Z. Qin, J. Weng, and K. Ren, "Privacy protection using a rechargeable battery for energy consumption in smart grids," *IEEE Netw.*, vol. 31, no. 1, pp. 59–63, Jan. 2017.

[6] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Cost-effective and privacy-preserving energy management for smart meters," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 486–495, Jan. 2015.

[7] A. M., "A novel non-cryptographic security services for advanced metering infrastructure in smart grid," *Commun. Appl. Electron.*, vol. 3, no. 7, pp. 35–39, Dec. 2015.

[8] S. Cleemput, M. A. Mustafa, E. Marin, and B. Preneel, "De-pseudonymization of smart metering data: Analysis and countermeasures," in *Proc. Global Internet Things Summit (GIoTS)*, Jun. 2018, pp. 1–6.

[9] N. Buescher, S. Boukoros, S. Bauregger, and S. Katzenbeisser, "Two is not enough: Privacy assessment of aggregation schemes in smart metering," *Proc. Privacy Enhancing Technol.*, vol. 2017, no. 4, pp. 198–214, Oct. 2017.

[10] M. A. Mustafa, S. Cleemput, A. Aly, and A. Abidin, "A secure and privacy-preserving protocol for smart metering operational data collection," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6481–6490, Nov. 2019.

[11] C. Thoma, T. Cui, and F. Franchetti, "Secure multiparty computation based privacy preserving smart metering system," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2012, pp. 1–6.

[12] D. Abbasinezhad-Mood and M. Nikooghadam, "An ultra-lightweight and secure scheme for communications of smart meters and neighborhood gateways by utilization of an ARM cortex-M microcontroller," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6194–6205, Nov. 2018.

[13] M. A. Mustafa, N. Zhang, G. Kalogridis, and Z. Fan, "DEP2SA: A decentralized efficient privacy-preserving and selective aggregation scheme in advanced metering infrastructure," *IEEE Access*, vol. 3, pp. 2828–2846, 2015.

[14] J. Ni, K. Zhang, X. Lin, and X. S. Shen, "Balancing security and efficiency for smart metering against misbehaving collectors," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1225–1236, Mar. 2019.

[15] L. Zhang, J. Zhang, and Y. H. Hu, "A privacy-preserving distributed smart metering temporal and spatial aggregation scheme," *IEEE Access*, vol. 7, pp. 28372–28382, 2019.

[16] Z. Guan, G. Si, J. Wu, L. Zhu, Z. Zhang, and Y. Ma, "Utility-privacy tradeoff based on random data obfuscation in internet of energy," *IEEE Access*, vol. 5, pp. 3250–3262, 2017.

[17] X. He, X. Zhang, and C.-C. J. Kuo, "A distortion-based approach to privacy-preserving metering in smart grids," *IEEE Access*, vol. 1, pp. 67–78, 2013.
[18] J. Bohli, C. Sorge, and O. Ugus, "A privacy model for smart metering," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2010, pp. 1–5.
[19] S. Tonyali, O. Cakmak, K. Akkaya, M. M. E. A. Mahmoud, and I. Guvenc, "Secure data obfuscation scheme to enable privacy-preserving state estimation in smart grid AMI networks," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 709–719, Oct. 2016.
[20] A. Beussink, K. Akkaya, I. F. Senturk, and M. M. E. A. Mahmoud, "Preserving consumer privacy on IEEE 802.11s-based smart grid AMI networks using data obfuscation," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 658–663.
[21] Y. Kim, E. C.-H. Ngai, and M. B. Srivastava, "Cooperative state estimation for preserving privacy of user behaviors in smart grid," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2011, pp. 178–183.
[22] H.-K. Wang, B.-C. Cheng, H. Chen, and P.-H. Hsu, "User daily behavior disturbance model to preserve privacy for smart metering," in *Proc. IEEE 5th Global Conf. Consum. Electron.*, Oct. 2016, pp. 1–2.
[23] P. Barbosa, A. Brito, and H. Almeida, "A technique to provide differential privacy for appliance usage in smart metering," *Inf. Sci.*, vols. 370–371, pp. 355–367, Nov. 2016.
[24] P. Barbosa, A. Brito, and H. Almeida, "Defending against load monitoring in smart metering data through noise addition," in *Proc. 30th Annu. ACM Symp. Appl. Comput.*, Apr. 2015, pp. 2218–2224.
[25] S. Wang, L. Cui, J. Que, D.-H. Choi, X. Jiang, S. Cheng, and L. Xie, "A randomized response model for privacy preserving smart metering," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1317–1324, Sep. 2012.
[26] P. Barbosa, A. Brito, H. Almeida, and S. Clauß, "Lightweight privacy for smart metering data by adding noise," in *Proc. 29th Annu. ACM Symp. Appl. Comput.*, Mar. 2014, pp. 531–538.
[27] A. S. Khwaja, A. Anpalagan, M. Naeem, and B. Venkatesh, "Smart meter data obfuscation using correlated noise," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7250–7264, Aug. 2020.
[28] G. Eibl and D. Engel, "Differential privacy for real smart metering data," *Comput. Sci. Res. Develop.*, vol. 32, nos. 1–2, pp. 173–182, Mar. 2017.
[29] Y. Kawamoto and T. Murakami, "Local distribution obfuscation via probability coupling," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2019, pp. 718–725.
[30] K. Mivule, "Utilizing noise addition for data privacy, an overview," 2013, *arXiv:1309.3958*.
[31] J. Kim and D. M. Jeong, "Truncated triangular distribution for multiplicative noise and domain estimation," *Sect. Government Statist. JSM*, vol. 2008, pp. 1023–1030, 2008.
[32] A. A. Ding, G. Miao, and S. S. Wu, "On the privacy and utility properties of triple matrix-masking," *J. Privacy Confidentiality*, vol. 10, no. 2, pp. 1–8, Jun. 2020.
[33] Y. Ma, Y.-X. Lin, and R. Sarathy, "The vulnerability of multiplicative noise protection to correlation-attacks on continuous microdata," *Sankhya B*, vol. 82, no. 2, pp. 305–327, Nov. 2020.
[34] Y.-X. Lin, L. Mazur, R. Sarathy, and K. Muralidhar, "Statistical information recovery from multivariate noise-multiplied data, a computational approach," *Trans. Data Priv.*, vol. 11, no. 1, pp. 23–45, 2018.
[35] J. J. Kim and W. E. Winkler, "Multiplicative noise for masking continuous data," Stat. Res. Division, U.S. Bur. Census, Washington, DC, USA, Tech. Rep. Statistics 2003-01, 2003.
[36] D. Mashima, A. Serikova, Y. Cheng, and B. Chen, "Towards quantitative evaluation of privacy protection schemes for electricity usage data sharing," *ICT Exp.*, vol. 4, no. 1, pp. 35–41, Mar. 2018.
[37] Y. Chen, J.-F. Martínez, P. Castillejo, and L. López, "A privacy-preserving noise addition data aggregation scheme for smart grid," *Energies*, vol. 11, no. 11, p. 2972, Nov. 2018.
[38] T. Zhang, Z. He, and R. B. Lee, "Privacy-preserving machine learning through data obfuscation," 2018, *arXiv:1807.01860*.
[39] K. Li, H. Pishro-Nik, and D. L. Goeckel, "Privacy against matching under anonymization and obfuscation in the Gaussian case," in *Proc. 52nd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2018, pp. 1–6.
[40] H. Cao, S. Liu, Z. Guan, L. Wu, H. Deng, and X. Du, "An efficient privacy-preserving algorithm based on randomized response in IoT-based smart grid," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, Oct. 2018, pp. 881–886.
[41] D. Yin and Q. Yang, "GANs based density distribution privacy-preservation on mobility data," *Secur. Commun. Netw.*, vol. 2018, pp. 1–13, Dec. 2018.
[42] K. Muralidhar and R. Sarathy, "Numerical data masking techniques for maintaining sub-domain characteristics," in *Proc. Joint UNECE/EUROSTAT Work Session Data Confidentiality UNECE/EUROSTAT Manchester*, 2007, pp. 1–10. [Online]. Available: https://ec.europa.eu/eurostat/documents/1001617/4569122/TOPIC-1-WP.05-IP-KRISH.pdf
[43] R. Masood, D. Vatsalan, M. Ikram, and M. A. Kaafar, "Incognito: A method for obfuscating web data," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 267–276.
[44] D. Tavella. (2019). *Mean Variance Product Random Variables*. Accessed: Feb. 6, 2021. [Online]. Available: https://www.researchgate.net/publication/332333452_Mean_and_Variance_of_the_Product_of_Random_Variables
[45] S. Aja-Fernandez and G. Vegas-Sanchez-Ferrero, *Statistical Analysis of Noise in MRI, Modeling, Filtering and Estimation*. Cham, Switzerland: Springer, 2016.
[46] A. Dytso, R. Bustin, H. V. Poor, and S. Shamai, "Analytical properties of generalized Gaussian distributions," *J. Stat. Distrib. Appl.*, vol. 5, no. 1, pp. 1–40, Dec. 2018.
[47] H. Pishro-Nik, *Introduction to Probability, Statistics, and Random Processes*. Kappa Research LLC, MA, USA, 2014. Accessed: Feb. 6, 2022. [Online]. Available: https://www.probabilitycourse.com/
[48] S. Ahn and J. A. Fessler, "Standard errors of mean, variance, and standard deviation estimators," Dept. Elect. Eng. Comput. Sci., Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. 413, Jul. 2003. Accessed: Feb. 6, 2022. [Online]. Available: http://web.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf
[49] Y. Li, J. Li, J. Qi, and L. Chen, "Robust cubature Kalman filter for dynamic state estimation of synchronous machines under unknown measurement noise statistics," *IEEE Access*, vol. 7, pp. 29139–29148, 2019.
[50] C. Walck. (2007). *Hand-book on Statistical Distributions for Experimentalists*. University of Stockholm. Stockholm, Sweden. Internal Report SUF?PFY/96?01. Accessed: Feb. 14, 2022. [Online]. Available: http://www.stat.rice.edu/~dobelman/textfiles/DistributionsHandbook.pdf

AHMED S. KHWAJA (Senior Member, IEEE) received the B.Sc. degree in electronic engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan, in 2001, and the M.Sc. and Ph.D. degrees in signal processing and telecommunications from the University of Rennes 1, Rennes, France, in 2004 and 2008, respectively. He is currently a Senior Research Associate with the WINCORE Laboratory, Ryerson University, Toronto, ON, Canada. His research interests include remote sensing, machine learning, and optimization problems in wireless communication systems and smart grid.

SERHAT ERKUCUK (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Middle East Technical University, Ankara, Turkey, in 2001, the M.Sc. degree in electrical and computer engineering from Ryerson University, Toronto, ON, Canada, in 2003, and the Ph.D. degree in engineering science from Simon Fraser University, Burnaby, BC, Canada, in 2007. In 2008, he was a NSERC Postdoctoral Fellow at The University of British Columbia, Vancouver, BC, Canada. He then joined Kadir Has University, Istanbul, Turkey, where he is currently a Full Professor. In 2018, he was a Visiting Professor at Ryerson University, where he conducted research on the design of small cells for 5G networks. His research interests include physical layer design of emerging communication systems, wireless sensor networks, and communication theory. He is a Marie Curie Fellow and a recipient of the Governor's General Gold Medal.

ALAGAN ANPALAGAN (Senior Member, IEEE) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto, Toronto, ON, Canada. He joined the Electrical and Computer Engineering Department, Ryerson University, Toronto, in 2001, and was promoted to a Full Professor, in 2010. He was with the department in administrative positions as the Associate Chair, the Program Director of electrical engineering, and the Graduate Program Director. He is a Registered Professional Engineer in ON, Canada; and a fellow of the Institution of Engineering and Technology and the Engineering Institute of Canada. He was a recipient of the IEEE Canada J. M. Ham Outstanding Engineering Educator Award, in 2018; the YSGS Outstanding Contribution to Graduate Education Award, in 2017; the Deans Teaching Award, in 2011; the Faculty Scholastic, Research and Creativity Award thrice from Ryerson University; the IEEE M. B. Broughton Central Canada Service Award, in 2016; the Exemplary Editor Award from IEEE Communications Society, in 2013; and the Editor-in-Chief Top 10 Choice Award in *Transactions on Emerging Telecommunications Technologies*, in 2012. He is a coauthor of a paper that received the IEEE SPS Young Author Best Paper Award, in 2015. He was the TPC Co-Chair of the IEEE Vehicular Technology Conference Fall 2017; the IEEE INFOCOM, in 2016; the IEEE GLOBECOM, in 2015; and the IEEE Personal Indoor Mobile Radio Communications, in 2011. He was the Vice Chair for the IEEE SIG on Green and Sustainable Networking and Computing with Cognition and Cooperation, from 2015 to 2018; the IEEE Canada Central Area Chair, from 2012 to 2014; the IEEE Toronto Section Chair, from 2006 to 2007; the Communications Society Toronto Chapter Chair, from 2004 to 2005; and the IEEE Canada Professional Activities Committee Chair, from 2009 to 2011. He was an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, from 2012 to 2014; IEEE COMMUNICATIONS LETTERS, from 2010 to 2013; and *EURASIP*, from 2004 to 2009. He was also a guest editor for six special issues published in IEEE, IET, and ACM.

BALA VENKATESH (Senior Member, IEEE) received the Ph.D. degree from Anna University, Chennai, India, in 2000. He is currently a Professor and the Academic Director of the Centre for Urban Energy, Ryerson University, Toronto, ON, Canada. His research interest includes power systems analysis and optimization.