MDMS: Efficient and Privacy-Preserving Multi-dimension and Multi-subset Data Collection for AMI Networks

Ahmad Alsharif, *Member, IEEE,* Mahmoud Nabil, *Member, IEEE,* Ahmed Sherif, *Member, IEEE,* Mohamed Mahmoud, *Member, IEEE,* and Min Song, *Fellow, IEEE*

***Abstract*—Advanced Metering Infrastructure (AMI) networks allow utility companies to collect fine-grained power consumption data of electricity consumers for load monitoring and energy management. This brings serious privacy concerns since the fine-grained power consumption data can expose consumers' activities. Privacy-preserving data aggregation techniques have been used to preserve consumers' privacy while allowing the utility to obtain only the consumers total consumption. However, most of the existing schemes do not consider the multi-dimensional nature of power consumption in which electricity consumption can be categorized based on the consumption type. They also do not consider multi-subset data collection in which the utility should be able to obtain the number of consumers whose consumption lies within a specific consumption range, and the overall consumption of each set of consumers. In this paper, we propose an efficient and privacy-preserving multi-dimensional and multi-subset data collection scheme, named "MDMS". In MDMS, the utility can obtain the total power consumption as well as the number of consumers of each subset in each dimension. In addition, for better scalability, MDMS allows the utility to delegate bill computation to the AMI networks' gateways using the encrypted readings and following dynamic prices in which electricity prices are different based on both the time and the consumption type. Moreover, MDMS uses lightweight operations in encryption, aggregation, and decryption resulting in low computation and communication overheads as given in our experimental results. Our security analysis demonstrates that MDMS is secure and can resist collusion attacks that aim to reveal the consumers' readings.**

***Index Terms*—Security, privacy preservation, multi-dimensional aggregation, multi-subset aggregation, smart grid, and AMI networks.**

I. Introduction

The smart grid integrates information and communication technologies into traditional power grids for improved robustness, efficiency, and reliability [1]. It provides two-way communications between the grid's entities to enable efficient and reliable power delivery to end consumers. Due to the two-way information flow, the smart grid can offer numerous advantages [2] to both electric utility companies and consumers, including but not limited to: (1) reducing the operation and management cost for utilities; (2) facilitating real-time load monitoring, energy management, and troubleshooting; (3) quicker restoration of electricity after power disturbances; and (4) lower electricity cost for consumers.

An Advanced metering infrastructure (AMI) network is a main component of the smart grid. In AMI networks, smart meters (SMs) deployed at consumers' houses are used to periodically report their nearly real-time power consumption readings to the utility at high rates, e.g., every few minutes. Then, the utility analyzes the data reported by the meters for real-time grid monitoring and energy management. For example, fine-grained power consumption analysis can be used to reduce the peak-to-average ratio which can lead to reducing electricity blackouts [3]. Also, power consumption analysis is required for real-time price-based demand/response programs in which electricity price changes depending on the supply-to-demand ratio especially during peak hours to balance energy supply and demand [4]. However, despite the importance of the fine-grained power consumption data collection, it poses potential threats to consumers' privacy since these data can reveal their daily activities. For example, a relatively low/high power consumption indicates the absence/presence of a consumer from/at his house. Also, non-intrusive load monitoring of a consumer's power consumption can reveal the appliances the consumer uses [5].

In order to preserve consumers' privacy, data aggregation techniques have been widely used in AMI networks [6]–[14]. Specifically, consumers send their fine-grained power consumption reading (PCR) to a local aggregator, called the gateway, which aggregates all the fine-grained PCRs and forwards an aggregated PCR to the utility for load monitoring and energy management. Therefore, the utility can only obtain the total power consumption of the consumers of an AMI network while hiding the fine-gained PCR of each consumer to preserve privacy.

Most of the existing data collection schemes [6]–[11] allow the utility to only obtain the total power consumption of the whole set of consumers in the AMI network. However, the utility needs more information to do more useful analysis on the data. In particular, power consumption data are *multi-dimensional* in nature, i.e., can be categorized based on the type of load, e.g., lamps, stove, oven, refrigerator, air heater/conditioning, and so on [15]. Furthermore, in addition

A. Alsharif is with the Department of Computer Science, University of Central Arkansas, Conway, AR, 72035 USA. E-mail: aalsharif@uca.edu.

M. Nabil is with the Department of Electrical & Computer Engineering, North Carolina A&T University, Greensboro, NC, 27401 USA. E-mail: mnmahmoud@ncat.edu.

A. Sherif is with the School of Computing Sciences and Computer Engineering, University of Southern Mississippi, Hattiesburg, MS 39406 USA. Email: ahmed.sherif@usm.edu.

M. Mahmoud is with the Department of Electrical & Computer Engineering, Tennessee Tech. University, Cookeville, TN 38505 USA. E-mail: mmahmoud@tntech.edu.

M. Song is with the Department of Electrical & Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 USA. Email: min.song@stevens.edu.

to learning the total power consumption of the whole set of consumers, the utility should also be able to learn the number of consumers whose PCs lie within a specific range, and the overall power consumption of each set of consumers, i.e., *multi-subset* data collection [16].

Multi-dimensional and multi-subset data collection can help in (1) efficient grid monitoring and energy management, (2) better prediction of power demands and developing direct load control demand/response programs [17] and (3) developing real-time billing based on dynamic pricing which is used for indirect load control demand/response programs [18]. Therefore, AMI networks need a data collection scheme that allows reporting PCRs in a multi-dimensional form and multi-subset aggregation based on different ranges for each dimension, i.e., a scheme that supports both multi-dimensional and multi-subset data aggregation simultaneously.

Therefore, we propose, in this paper, an efficient and privacy-preserving **m**ulti-**d**imensional **m**ulti-**s**ubset data collection scheme for AMI network, named "MDMS". Based on, but not limited to, the secure computations over encrypted data using the k-nearest neighbor (kNN) similarity measurement, we develop MDMS that allows data aggregation and the computation of electricity bills in the ciphertext domain. In MDMS, each smart meter reports multi-dimension and multi-subset power consumption in an encrypted data vector to the gateway which aggregates all encrypted vectors and sends a single aggregated vector to the utility. Upon decryption of that vector, the utility can obtain (1) the total electricity consumption of each subset in each dimension and (2) the number of consumers of each subset in each dimension.

The novelty and contributions of this paper can be summarized as follows.

1) We propose a privacy-preserving multi-dimensional and multi-subset data collection scheme for AMI networks. MDMS can achieve multi-dimensional data aggregation with multi-subsets for each dimension whereas closely similar schemes can aggregate either multi-dimensional or multi-subset data but not both simultaneously.
2) MDMS utilizes masking technique such that it can resist collusion attacks launched by the utility, gateways, smart meters and external adversaries to reveal the power consumption data of the individual consumers. Masks are computed efficiently and in an offline way.
3) Compared to the proposed schemes in [12]–[14], MDMS uses lightweight operations in encryption, aggregation and decryption which results in better performance in terms of communication and computation overheads.
4) MDMS enables the utility to delegate the computation of the electricity bills to the gateways for better scalability, without exposing its private key to the gateways and without violating consumers' privacy. The bills are computed using the consumers' fine-grained encrypted PCRs and following dynamic pricing in which electricity prices are different based on the billing time and the load type.

A preliminary version of this paper has been published in [19]. The main difference between [19] and this paper are as follows. First, [19] supports only multi-dimensional data aggregation while this paper considers multi-subset multi-dimensional data aggregation. Second, [19] does not address collusion attacks and dynamic billing which are addressed in this paper. Third, extensive analysis and simulation have been added to this paper. These include a comprehensive security analysis, and updated simulation results.

The remainder of this paper is organized as follows. Related works are discussed in Section II. The considered system models and design requirements are presented in Section III. Preliminaries are given in Section IV. The proposed data collection scheme is explained in Section V The security and privacy preservation analysis and performance evaluations are given in Sections VI and VII, respectively. Conclusions are drawn in Section VIII.

II. Related Works

Data aggregation techniques have been widely used to preserve consumers' privacy in AMI networks [6]–[11]. The schemes proposed in [6]–[8] use one-time masking such that each meter masks its PCR and send the masked PCR to the gateways. In this way, gateways cannot access the individual PCRs to preserve consumers' privacy. When the all the masked PCRs are aggregated, all the masks cancel each other and thus the utility can obtain only the total power consumption for the consumers of the AMI network.

On the other hand, the schemes proposed in [9]–[11] preserve consumers' privacy by exploiting the additive homomorphic property of the Paillier cryptosystem. In specific, each meter encrypts its PCR using the Paillier cryptosystem and sends the encrypted PCR to the gateways. All the encrypted PCRs are aggregated in the ciphertext domain, i.e., gateways cannot decrypt them and thus cannot violate consumers' privacy. Finally, the utility can decrypt the aggregated ciphertext to recover the total consumption of the consumers of the AMI network. However, the above data collection schemes are designed to collect the total consumption of the AMI network's consumers and cannot handle either the multi-dimensional nature of power consumption or multi-subset data aggregation.

In the literature, few schemes have addressed either multi-dimensional data aggregation [12] or multi-subset data aggregation [13], [14] but not both of them simultaneously. In [12], Lu *et al.* proposed the first attempt to realize multi-dimensional power consumption data collection by using a super-increasing sequence to represent the multi-dimensional power consumption and then encrypt it using Paillier cryptosystem. Then, all meters' ciphertexts are aggregated into a single ciphertext that is sent to the utility. The utility can perform the decryption process to obtain the total power consumption of the consumers for each data dimension. However, the scheme proposed in [12] is limited to multi-dimensional data collection for the whole set of consumers and cannot be used efficiently for multi-subset data aggregation within each dimension. In addition, it does not address the possible collusion between gateways and the utility to violate consumers' privacy.

In [13], Lu *et al.* have made the first attempt to develop subset aggregation scheme based on the additive homomorphic properties of composite order cryptographic groups [20]. However, [13] is limited to two subset data aggregation only and

TABLE I
COMPARISON BETWEEN MDMS AND RELATED SCHEMES.

	MDMS	[12]	[13]	[14]
Privacy Preservation	√	√	√	√
Collusion Resistance	√	NA	√	√
Aggregation				
Multi-dimensional only	√	√	×	√‡
Multi-subset only	√	×	√*	√‡
Multi-subset for each dimension	√	×	×	×
Billing				
Basic Billing	√	NA	NA	√Δ
Advanced Billing	√	NA	NA	×
Low communication computation overhead	√	×	×	×

NA: Not Addressed
*: Cannot support more than two subsets.
Δ: Unlike MDMS, [14] involves a trusted third party to compute bills.
‡: Can support either multi-dimensional or multi-subset aggregation but not multi-subset for each dimension.

cannot handle either the multi-dimensional nature of power consumption or the case of more than two subsets. Moreover, the decryption process uses Pollard's lambda algorithm which becomes inefficient as the size of the message to be decrypted increases. Similar to [12], Li *et al.* in [14] used two super-increasing sequences and Paillier cryptosystem to realize privacy-preserving multi-subset aggregation. Although [14] can support either multi-subset or multi-dimensional data aggregation, it cannot handle multi-subset data aggregation within each data dimension. In addition, [12], [14] are not efficient in terms of communication and computation since they utilize the Paillier cryptosystem which typically requires long times for encryption and decryption, as will be shown in Section VII.

In Tab. I, we summarize the comparison of MDMS against the closely similar schemes [12]–[14].

III. SYSTEM MODELS AND DESIGN REQUIREMENTS

In this section, we describe the considered network and attack models. Also, we give the design requirements.

A. Network Model

As shown in Fig. 1, our network model consists of the following entities, a key distribution center, the utility, gateways, and a set of smart meters forming an AMI networks. The role of each entity is described below.

- *Key Distribution Center (KDC).* The KDC is responsible for generating and distributing the public parameters and secret keys to the smart meters and the utility. After key generation and distribution, the KDC does not participate in the periodic data collection or billing processes.
- *The Utility*. The utility is responsible for the power consumption data analysis to monitor the load and manage the electricity generation to meet the dynamic demand. Therefore, it needs to collect the power consumption data of consumers periodically.
- *The Gateway (GW)*. Each AMI network has a GW that connects the meters of the network to the utility. An AMI network can cover a neighborhood. For the periodic data collection process, the GW collects and aggregates the encrypted reports received from the meters and forwards the encrypted aggregated data to the utility. For billing purposes, each GW computes electricity bills for the consumers without violating their privacy.
- *SMs*. We consider that each AMI network has a set of smart meters $\mathbb{SM} = \{\text{SM}_i, 1 \leq i \leq |\mathbb{SM}|\}$. A smart meter is installed at each consumer's house to periodically report encrypted multi-dimensional and multi-subset power consumption to the utility via the GW.

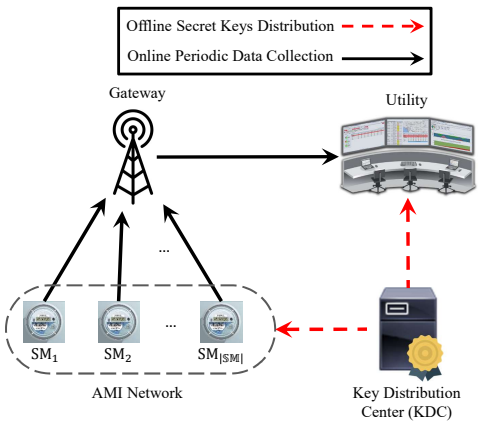


Fig. 1. The considered AMI network model.

MDMS not only collects power consumption data in a multi-dimensional form, but also allows multi-subset aggregation based on different ranges for each dimension. For the multi-dimensional nature of PC, we consider a set of m-dimensions that represent m-types of power consumption data $\{D_j : 1 \leq j \leq m\}$ to be reported. For each type of PC, we assume that there are n-subsets equivalent to n consecutive power consumption ranges, i.e., $\Big\{[0, R_1], (R_1, R_2] \ldots, (R_{n-1}, R_n]\Big\}$, where $\{R_1, \ldots, R_n\}$ are the subsets' limits. If the PCR (R) satisfies $R_{k-1} < R \leq R_k$ then, it is reported within subset k. Let $\mathbb{SM}_k$ be the set of smart meters whose reported PCRs lie in subset k and $|\mathbb{SM}_k|$ be the size of this subset, where $\mathbb{SM}_k \subseteq \mathbb{SM}$. The sum of all subsets' sizes is equal to the total number of smart meters, i.e., $\sum_{k=1}^{n} |\mathbb{SM}_k| = |\mathbb{SM}|$, and the PCR reported by a meter cannot belong to two different subsets, i.e., $\mathbb{SM}_k \bigcap \mathbb{SM}_\ell = \phi$ for any $k \neq \ell$.

B. Attack Model

The GW and the utility are honest-but-curious, i.e., although they follow the proposed scheme correctly, they attempt to learn the individual PCRs of the consumers. In specific, the GW receives the encrypted PCRs and may try to infer any information about the PCR of any consumer. Also, the utility may use its secret key to reveal the PCR of any consumer. The smart meters are also honest-but-curious. Each smart meter reports correct data to the utility, however, it also tries to learn the individual PCRs of other meters. Moreover, there exist external adversaries $\mathscr{A}$ that eavesdrop the communications between the different entities in the network aiming to obtain

individual meter's readings. Furthermore, attackers can work individually or collude with each other to launch stronger attacks. In a collusion attack an adversary is considered to control the utility, the GW and a set of smart meters of size $|\mathbb{SM}_{\mathscr{A}}| < |\mathbb{SM}| - 1$ in order to revel the PCR of an honest meter. We focus in this paper on preserving consumers' privacy and resisting the collusion between different entities that aim to obtain individual SMs' readings. Other attacks are beyond the scope of the paper.

C. Design Requirements

Based on the multi-dimensional and multi-subset data collection objective and the aforementioned threat model, the following functional and privacy requirements should be met.

1) Functional Requirements:

(F1) At each PCR reporting period, MDMS should allow the utility to efficiently obtain the total electricity consumption and the number of consumers "size" of each subset in each dimension.

(F2) MDMS should allow electricity bill computation based on fluctuating electricity prices. Two billing cases should be supported by MDMS: basic billing, in which electricity price within the same billing interval is the same regardless of the type of the power consumption load, and advanced billing, in which electricity prices are different based on the type of the PC's load.

2) Security and Privacy Requirements: At each PCR reporting period, the following requirements should be met.

(SP1) Consumers' privacy preservation. No entity should be able to access the individual PCR of any consumer at any time.

(SP2) Aggregated data confidentiality. No entity, except the utility, should be able to access the aggregated power consumption and the number of consumers "size" of each subset within each dimension.

(SP3) Collusion resistance. MDMS should resist collusion attacks in which attackers may collude to obtain the PCRs of the individual consumers.

IV. Preliminaries

A. Bilinear Pairings

Let $\mathbb{G}_1$ be an additive cyclic group, $\mathbb{G}_2$ be a multiplicative cyclic group of the same prime order q, and P be a generator of $\mathbb{G}_1$. A pairing $\hat{e}$:$\mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$ has the following properties.

- *Bilinearity*: $\hat{e}(aP_1, bP_2) = \hat{e}(P_1, P_2)^{ab} \in \mathbb{G}_2 \ \forall \ P_1, P_2 \in \mathbb{G}_1$ and $a, b \in \mathbb{Z}_q^*$.
- *Non-degeneracy*: $\hat{e}(P, P) \neq 1_{\mathbb{G}_2}$.

B. Non-interactive key establishment

ID-based cryptography allows any two entities, such as SM_i and SM_ℓ, to establish a static key in a non-interactive manner [22]. In specific, SM_i uses its private key X_i along with the public key Q_ℓ of another meter SM_ℓ to compute $K_{i\ell} = \hat{e}(X_i, Q_\ell) = \hat{e}(sQ_i, Q_\ell) = \hat{e}(Q_i, Q_\ell)^s$. Similarly, SM_ℓ uses its private key SM_ℓ along with SM_i's public key Q_i to compute the same key as $K_{i\ell} = \hat{e}(Q_i, X_\ell) = \hat{e}(Q_i, sQ_\ell) = \hat{e}(Q_i, Q_\ell)^s$.

TABLE II
Main notations

Notation	Description
$q, \mathbb{G}_1, \mathbb{G}_2, \hat{e}, P, Q, H_1, H_2, H_3$	Public parameters for the ID-based signature scheme [21]
$\mathcal{H}(K, m)$	Keyed hash function, K is the key, m is the message
$\mathbb{SM}$	Set of smart meters $\mathbb{SM} = \{\text{SM}_i, 1 \leq i \leq \|\mathbb{SM}\|\}$
SM_i/ID_i	i-th smart meter / Identity of SM_i
$X_{i,0}, X_{i,1}$	ID-based private keys of SM_i
$Q_{i,0}, Q_{i,1}$	ID-based public keys of SM_i
$\mathcal{MK}$	Master kNN key set for the utility $\mathcal{MK} = \{\mathbf{M}_1, \mathbf{M}_2, \mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3, \mathbf{N}_4\}$
$\mathcal{K}_i$	kNN key for SM_i derived from $\mathcal{MK}$ $\mathcal{K}_i = \{\mathbf{A}_i\mathbf{N}_1,\ \mathbf{B}_i\mathbf{N}_2,\ \mathbf{C}_i\mathbf{N}_3,\ \mathbf{D}_i\mathbf{N}_4\}$
j	j-th dimension of power consumption $1 \leq j \leq m$
t	reporting period
$R_{it}^{(j)}$	PCR of SM_i for dimension j at reporting period t
$\mathbf{r}_{it}$	Reading vector of SM_i at reporting period t
$\mathbf{m}_{it}$	Masking vector used by SM_i at reporting period t
$\mathbf{w}_{it}$	The masked vector of SM_i at reporting period t $\mathbf{w}_{it} = \mathbf{r}_{it} + \mathbf{m}_{it}$
$\mathbf{c}_{it}$	kNN ciphertext computed by SM_i for $\mathbf{w}_{it}$
S_i, Y_i	ID-based signature components computed by SM_i
$\mathbf{c}_{at}$	Aggregated ciphertext computed by the GW
$\mathbf{m}_{at}$	Aggregated masks recovered by the utility $\mathbf{m}_{at} = \sum_{i=1}^{\|\mathbb{SM}\|} \mathbf{m}_{it} = 0$
$\mathbf{r}_{at}$	Aggregated reading vector recovered by the utility
$\mathbf{w}_{at}$	Aggregated masked vector recovered by the utility $\mathbf{w}_{at} = \mathbf{r}_{at} + \mathbf{m}_{at}$
$\mathbf{r}_i$	Consumption vector of SM_i during a billing interval
$\mathbf{p}$	Pricing vector within a billing interval

C. Secure k-nearest neighbor computation

Secure computations over encrypted data using the k-nearest neighbor (kNN) similarity measurement has been widely used in several applications such as keyword searching [23]–[26], multi-recipient AMI networks [27], and location-based applications [28], [29]. Based on, but not limited to, the kNN, we develop MDMS that allows data aggregation and the computation of electricity bills in the ciphertext domain .

V. The MDMS Scheme

In this section, we give the details of MDMS. For better readability, we define the main notations that will be used in next subsections in Tab. II. We use lowercase bold notation for vectors and uppercase bold notation for matrices. For example, $\mathbf{r}$ and $\mathbf{m}$ are vectors while $\mathbf{M}$ and $\mathbf{N}$ are matrices.

A. Overview

Fig. 2 shows the structure of the reading vector, $\mathbf{r}_{it}$, used in MDMS. As shown in the figure, $\mathbf{r}_{it}$ is constructed of m blocks representing m data dimensions. For each dimension D_j, n elements are used as subset indicators and another n elements are used to report the PCR of this dimension. For instance, let $R_{it}^{(j)}$ be the PCR of SM_i for dimension D_j at reporting period t. As shown in Fig. 2, if $R_{it}^{(j)}$ lies within subset k,

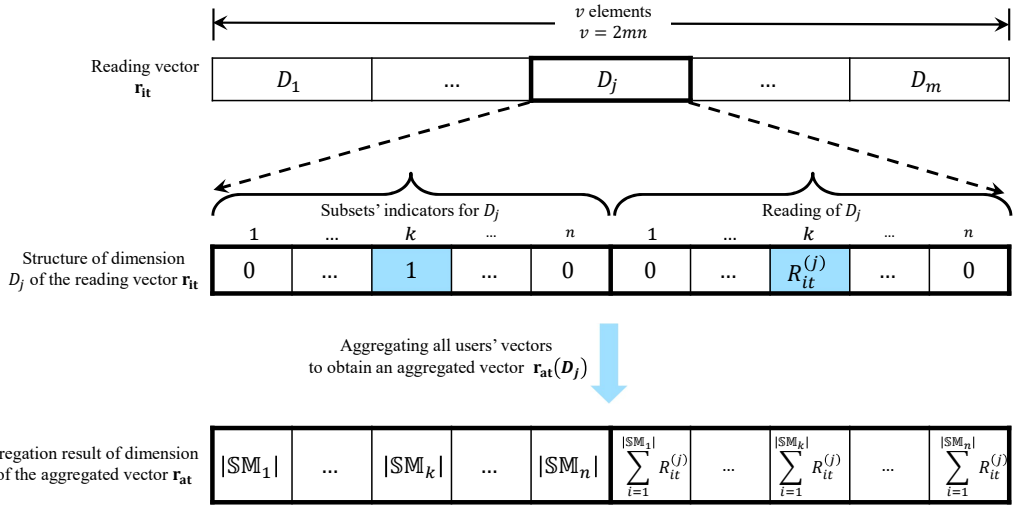


Fig. 2. Reading vector structure and the aggregation result.

then SM_i reports one at the k-th subset indicator and $R_{it}^{(j)}$ at the k-th reading location, and set all other elements to zeros. Then, SM_i masks $\mathbf{r}_{it}$ by a masking vector $\mathbf{m}_{it}$ that can be pre-computed as will be explained later in this section. The masked vector, $\mathbf{w}_{it} = \mathbf{r}_{it} + \mathbf{m}_{it}$, is then encrypted using the kNN encryption technique to generate a ciphertext vector $\mathbf{c}_{it}$ to be sent to the GW. The GW aggregates all the encrypted vectors received from all meters and outputs an aggregated vector, $\mathbf{c}_{at}$, which is the encryption of the aggregated masked vector $\mathbf{w}_{at} = \mathbf{r}_{at} + \mathbf{m}_{at} = \sum \mathbf{r}_{it} + \sum \mathbf{m}_{it}$. After aggregation, the GW sends the utility the aggregated ciphertext $\mathbf{c}_{at}$. The masks are generated such that they cancel each other when all meters' vectors are aggregated, i.e., $\mathbf{m}_{at} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{m}_{it} = 0$. Therefore, when the utility decrypts $\mathbf{c}_{at}$, it can obtain the aggregated vector $\mathbf{w}_{at} = \mathbf{r}_{at} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{r}_{it}$. As shown in Fig. 2, the content of $\mathbf{r}_{at}$ for each dimension is the size of each subset $|\mathbb{SM}_k|$ and the total power consumption of each subset $\sum_{i=1}^{|\mathbb{SM}_k|} R_{it}^{(j)}$ at the reporting period t.

B. System Initialization

System initialization, carried out by the KDC, consists of the following phases (1) generation of public system parameters, (2) generation of ID-based public/private key pairs, and (3) generation of kNN meters' keys and utility key.

1) Generation of public system parameter: The KDC generates the public parameters as follows. It

1) generates the bilinear pairing parameters $(\mathbb{G}_1, \mathbb{G}_2, \hat{e}, P, q)$;
2) chooses a random element $s \in \mathbb{Z}_q^*$ and computes $Q = sP \in \mathbb{G}_1$;
3) chooses three cryptographic hash functions H_1, H_2, H_3 defined as $H_1, H_2 : \{0,1\}^* \rightarrow \mathbb{G}_1$ and $H_3 : \{0,1\}^* \rightarrow \mathbb{Z}_q^*$;
4) chooses a keyed hash function $\mathcal{H}(K, m)$ where K is the key used to compute a keyed-hash on a message m.

Then, s is kept as a master secret and the public system parameters $param = \{\mathbb{G}_1, \mathbb{G}_2, \hat{e}, P, Q, H_1, H_2, H_3, \mathcal{H}\}$ are published.

2) Generation of ID-based public/private key pairs: Each SM_i with an identity ID_i receives from the KDC two private keys $X_{i,0} = sQ_{i,0}$ and $X_{i,1} = sQ_{i,1}$, where $Q_{i,0} = H_1(\text{ID}_i, 0)$ and $Q_{i,1} = H_1(\text{ID}_i, 1)$ are the corresponding public keys. Similarly, the GW receives from the KDC private keys $X_{g,0} = sQ_{g,0}$ and $X_{g,1} = sQ_{g,1}$ and the corresponding public keys $Q_{g,0} = H_1(\text{ID}_g, 0)$ and $Q_{g,1} = H_1(\text{ID}_g, 1)$. Since theses public keys are ID-based, the meters and the GW do not need to obtain digital certificates from the KDC to certify the public keys.

3) Generation of kNN keys: The KDC generates a random binary v-dimensional vector $\mathbf{s}$ to be used as a splitting indicator for the kNN encryption technique. The size of $\mathbf{s}$ is $v = 2mn$. Then, the KDC generates a master key set, $\mathcal{MK} = \{\mathbf{M}_1, \mathbf{M}_2, \mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3, \mathbf{N}_4\}$, where each element in $\mathcal{MK}$ is a $v \times v$ invertible random matrix. $\mathcal{MK}$ is sent to the utility via a secure channel. Also $\mathcal{MK}$ is used to drive a unique key for each meter.

Generation of meters' keys. For each SM_i, the KDC uses $\mathcal{MK}$ to generate a unique key $\mathcal{K}_i$ as $\mathcal{K}_i = \{\mathbf{A}_i\mathbf{N}_1, \mathbf{B}_i\mathbf{N}_2, \mathbf{C}_i\mathbf{N}_3, \mathbf{D}_i\mathbf{N}_4\}$, where $\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i$ and $\mathbf{D}_i$ are $v \times v$ invertible random matrices such that $\mathbf{A}_i + \mathbf{B}_i = \mathbf{M}_1$ and $\mathbf{C}_i + \mathbf{D}_i = \mathbf{M}_2$. Finally, the KDC sends $\mathcal{K}_i$ to SM_i via a secure channel.

C. Smart meter: PCR Encryption

For each reporting period t, each $\text{SM}_i \in \mathbb{SM}$ builds a data vector, $\mathbf{r}_{it}$, as shown in Fig. 2, that contains m PCRs $\left(R_{it}^{(j)} : j \in \{1, 2, \ldots, m\}\right)$ where $R_{it}^{(j)}$ is the PCR of SM_i for dimension j at reporting period t. Then, SM_i encrypts this vector and report the encrypted vector to the GW by performing the following six steps.

- Step 1: SM_i builds a v-dimensional data vector $\mathbf{r}_{it}$ that contains its m PCRs as shown in Fig. 2.
- Step 2: SM_i builds a v-dimensional masking vector $\mathbf{m}_{it}$ where the z-th element of $\mathbf{m}_{it}$ is computed as follows

$$\mathbf{m}_{it}(z) = \sum_{\substack{\ell < i \\ 1 \le \ell \le |\mathbb{SM}|}} \mathcal{H}(K_{i\ell}\ ,\ t \parallel z) - \sum_{\substack{\ell > i \\ 1 \le \ell \le |\mathbb{SM}|}} \mathcal{H}(K_{i\ell}\ ,\ t \parallel z)$$

where $\mathcal{H}$ is the public keyed hash function and $K_{i\ell}$ is the symmetric key shared between SM_i and SM_ℓ computed in a non-interactive way as explained in subsection IV-B. It should be noted that the masking vector $\mathbf{m}_{it}$ can be computed offline, i.e, before the reporting period starts.

- Step 3: SM_i masks the data vector $\mathbf{r}_{it}$ using the masking vector $\mathbf{m}_{it}$ to generate a vector $\mathbf{w}_{it}$ as $\mathbf{w}_{it} = \mathbf{r}_{it} + \mathbf{m}_{it}$.
- Step 4: SM_i splits the vector $\mathbf{w}_{it}$ into two random v-dimension vectors $\mathbf{w}'_{it}$ and $\mathbf{w}''_{it}$ using the splitting indicator $\mathbf{s}$. For the z-th element in the vector $\mathbf{w}_{it}$, splitting is done as follows

$$\begin{aligned} \mathbf{w}'_{it}(z) = \mathbf{w}''_{it}(z) = \mathbf{w}_{it}(z) \quad \text{if} \quad \mathbf{s}(z) = 1 \\ \mathbf{w}'_{it}(z) = r, \mathbf{w}''_{it}(z) = \mathbf{w}_{it}(z) - \mathbf{w}'_{it}(z) \quad \text{if} \quad \mathbf{s}(z) = 0 \end{aligned}$$

where r is a random number.

- Step 5: SM_i uses $\mathbf{w}'_{it}$, $\mathbf{w}''_{it}$ and its secret key $\mathcal{K}_i$ to compute a ciphertext $\mathbf{c}_{it}$ as

$$\mathbf{c}_{it} = \{\mathbf{w}'_{it}\mathbf{A}_i\mathbf{N}_1 \ , \ \mathbf{w}'_{it}\mathbf{B}_i\mathbf{N}_2 \ , \ \mathbf{w}''_{it}\mathbf{C}_i\mathbf{N}_3 \ , \ \mathbf{w}''_{it}\mathbf{D}_i\mathbf{N}_4\}$$

where $\mathbf{c}_{it}$ is a row vector of size $1 \times 4v$.

- Step 6: SM_i uses its private keys $X_{i,0}$ and $X_{i,1}$ to generate a signature on $\mathbf{c}_{it}$ by using the signature scheme proposed in [21]. First SM_i computes $P_t = H_2(t)$ and $h_i = H_3(\mathbf{c}_{it}, \text{ID}_i, t)$. Then, SM_i chooses a random element $y_i \in \mathbb{Z}_q^*$. Finally the signature components Y_i and S_i are computed as follows

$$Y_i = y_i P$$

$$S_i = y_i P_t + X_{i,0} + h_i X_{i,1}$$

- Step 7: SM_i sends to the GW the following report.

$$\mathbf{c}_{it} \parallel \text{ID}_i \parallel t \parallel S_i \parallel Y_i$$

D. Gateway: Efficient Aggregation

After collecting all the meters' reports, the GW verifies the received signatures, aggregate all the ciphertexts and send the aggregated message to the utility by performing the following steps

- Step 1: The GW computes $h'_i = H_3(\mathbf{c}_{it}, \text{ID}_i, t)$ for $1 \leq i \leq |\mathbb{SM}|$.
- Step 2: The GW verifies the received signatures to ensure reports' integrity and the authenticity of reports' senders. Efficient batch verification process can be done by checking

$$\hat{e}\left(\sum_{i=1}^{|\mathbb{SM}|} S_i, P\right) \stackrel{?}{=} \hat{e}\left(\sum_{i=1}^{|\mathbb{SM}|} Y_i, P_t\right) \hat{e}\left(Q, \sum_{i=1}^{|\mathbb{SM}|} Q_{i,0} + h'_i Q_{i,1}\right)$$

- Step 3: The GW computes the aggregated ciphertext $\mathbf{c}_{at}$ as

$$\begin{aligned} \mathbf{c}_{at} &= \sum_{i=1}^{|\mathbb{SM}|} \mathbf{c}_{it} \\ &= \begin{Bmatrix} \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{A}_i\mathbf{N}_1 & , & \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{B}_i\mathbf{N}_2 \\ \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}''_{it}\mathbf{C}_i\mathbf{N}_3 & , & \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}''_{it}\mathbf{D}_i\mathbf{N}_4 \end{Bmatrix} \\ &= \{\mathbf{c}_{at,1}, \mathbf{c}_{at,2}, \mathbf{c}_{at,3}, \mathbf{c}_{at,4}\} \end{aligned}$$

where $\mathbf{c}_{at}$ consists of four components $\{\mathbf{c}_{at,1}, \mathbf{c}_{at,2}, \mathbf{c}_{at,3}, \mathbf{c}_{at,4}\}$ and each component is row vector of size v.

- Step 4: The GW uses its private keys $X_{g,0}$ and $X_{g,1}$ to generate a signature on $\mathbf{c}_{at}$. First the GW computes $h_g = H_3(\mathbf{c}_{at}, \text{ID}_g, t)$. Then, it chooses a random element $y_g \in \mathbb{Z}_q^*$. Finally the signature components Y_g and S_g are computed as follows

$$Y_g = y_g P$$

$$S_g = y_g P_t + X_{g,0} + h_g X_{g,1}$$

- Step 5: The GW sends to the utility the following report

$$\mathbf{c}_{at} \parallel \text{ID}_g \parallel t \parallel S_g \parallel Y_g$$

E. Utility: Decryption and aggregated data recovery

After receiving $\mathbf{c}_{at} \parallel \text{ID}_g \parallel t \parallel S_g \parallel Y_g$ from the GW, the utility performs the following steps.

- Step 1: The utility computes $h'_g = H_3(\mathbf{c}_{at}, \text{ID}_g, t)$.
- Step 2: The utility verifies the signature by checking $\hat{e}\left(S_g, P\right) \stackrel{?}{=} \hat{e}\left(Y_g, P_t\right) \hat{e}\left(Q, Q_{g,0} + h'_g Q_{g,1}\right)$
- Step 3: The utility uses its key $\mathcal{MK}$ to recover $\mathbf{w}'_{at} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}$ from the first two components of $\mathbf{c}_{at}$ as follows

$$\mathbf{w}'_{at} = \mathbf{c}_{at,1}\mathbf{N}_1^{-1}\mathbf{M}_1^{-1} + \mathbf{c}_{at,2}\mathbf{N}_2^{-1}\mathbf{M}_1^{-1} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}$$

The correctness of this equation is as follows.

$$\begin{aligned} \mathbf{w}'_{at} &= \mathbf{c}_{at,1}\mathbf{N}_1^{-1}\mathbf{M}_1^{-1} + \mathbf{c}_{at,2}\mathbf{N}_2^{-1}\mathbf{M}_1^{-1} \\ &= \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{A}_i\mathbf{N}_1\mathbf{N}_1^{-1}\mathbf{M}_1^{-1} + \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{B}_i\mathbf{N}_2\mathbf{N}_2^{-1}\mathbf{M}_1^{-1} \\ &= \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{A}_i\mathbf{M}_1^{-1} + \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{B}_i\mathbf{M}_1^{-1} \\ &= \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\left(\mathbf{A}_i + \mathbf{B}_i\right)\mathbf{M}_1^{-1} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it}\mathbf{M}_1\mathbf{M}_1^{-1} \\ &= \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}'_{it} \end{aligned}$$

- Step 4: In a similar process, the utility recovers $\mathbf{w}''_{at} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}''_{it}$ from the last two components of $\mathbf{c}_{at}$ as follows

$$\mathbf{w}''_{at} = \mathbf{c}_{at,3}\mathbf{N}_3^{-1}\mathbf{M}_2^{-1} + \mathbf{c}_{at,4}\mathbf{N}_4^{-1}\mathbf{M}_2^{-1} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{w}''_{it}$$

- Step 5: The utility uses the vector $\mathbf{s}$ to merge $\mathbf{w}'_{at}$ and $\mathbf{w}''_{at}$ to obtain $\mathbf{w}_{at}$ as follws

$$\begin{aligned} \mathbf{w}_{at}(z) = \mathbf{w}'_{at}(z) \quad \text{if} \quad \mathbf{s}(z) = 1 \\ \mathbf{w}_{at}(z) = \mathbf{w}'_{at}(z) + \mathbf{w}''_{at}(z) \quad \text{if} \quad \mathbf{s}(z) = 0 \end{aligned}$$

The result of the decryption process is $\mathbf{w}_{at} = \mathbf{r}_{at} + \mathbf{m}_{at} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{r}_{it} + \sum_{i=1}^{|\mathbb{SM}|} \mathbf{m}_{it}$. However, since the masks are generated such that $\sum_{i=1}^{|\mathbb{SM}|} \mathbf{m}_{it} = 0$, the decryption result is the vector $\mathbf{w}_{at} = \mathbf{r}_{at} = \sum_{i=1}^{|\mathbb{SM}|} \mathbf{r}_{it}$. As discussed in subsection V-A, the contents of the vector $\mathbf{r}_{at}$ for each dimension are the size

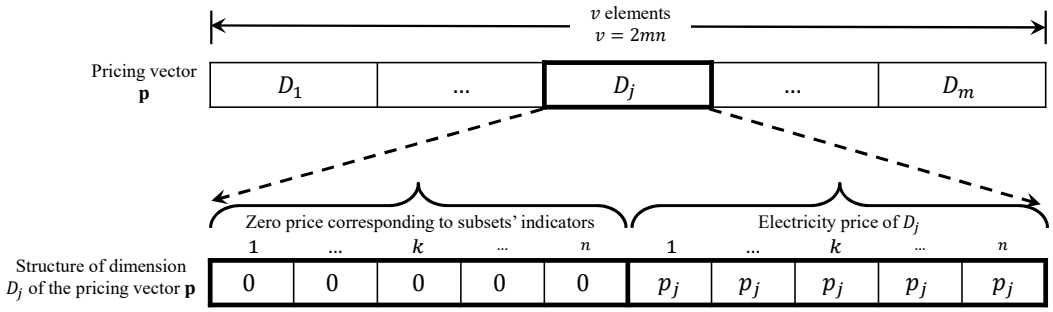


Fig. 3. Price vector structure.

of each subset $|\mathbb{SM}_k|$ and the total power consumption of each subset $\sum_{i=1}^{|\mathbb{SM}_k|} R_{it}^{(j)}$. Therefore, MDMS achieves the functional requirement (F1).

F. Billing based on Dynamic Pricing

Assume that each day is divided into several billing intervals. The utility can generate electricity bills for customers at the end of each billing interval. However, to consider the network scalability, the utility can delegate the bill computation to GWs, i.e., each local GW can compute the electricity bills of its consumers without exposing the utility's private key to the GWs and without violating consumers' privacy, instead of making the utility compute the bills for all the consumers in the system.

First, based on the time and the total power consumption of consumers for each dimension, the utility can determine the electricity prices for each type of consumption, i.e., for m power consumption types, the utility can set electricity prices to $\{p_1, \ldots, p_m\}$ for the m types for the advanced billing case. Then, in order to compute the electricity bill, the meters, the GW, and the utility, perform the following computations.

1) Smart meters: Assume that each SM_i reports T power consumption reports during each billing interval, i.e., $1 \leq t \leq T$. At the end of each billing interval, (report sent at $t = T$), in order to encrypt its data vector $\mathbf{r}_{iT}$, SM_i follows the same steps explained in subsection V-C except that SM_i uses a masking vector $\mathbf{m}_{iT}$ that is equal to the negative summation of all the previous $T-1$ masks used.

2) Utility: The utility delegates the bill computation to the GW without violating consumer's privacy or revealing its secret key set $\mathcal{MK}$. Therefore, the utility sends the GW an encrypted price vector such that the GW can only compute the bill for each user without being able to decrypt their encrypted reports. The encrypted price vector is generated as follows.

- Step 1: The utility builds a v-dimensional price vector $\mathbf{p}$ that contains the m prices as shown in Fig. 3. As shown in the figure, for each dimension D_j, the utility sets the locations corresponding to subset indicators to zeros and sets the locations corresponding to the PCR to the electricity price of that dimension. The price vector $\mathbf{p}$ is a column vector of size $v \times 1$.
- Step 2: The utility splits the price vector $\mathbf{p}$ into two random v-dimension column vectors $\mathbf{p}'$ and $\mathbf{p}''$ using the splitting indicator $\mathbf{s}$. For the z-th element in the vector $\mathbf{p}$, splitting is done as follows.

$$\begin{aligned} \mathbf{p}'(z) = \mathbf{p}''(z) = \mathbf{p}(z) \quad &\text{if} \quad \mathbf{s}(z) = 0 \\ \mathbf{p}'(z) = r, \mathbf{p}''(z) = \mathbf{p}(z) - \mathbf{p}'(z) \quad &\text{if} \quad \mathbf{s}(z) = 1 \end{aligned}$$

where r is a random number.
- Step 3: The utility uses $\mathbf{p}'$, $\mathbf{p}''$ and its secret key $\mathcal{MK}$ to compute a ciphertext $\mathbf{c}_u$ as

$$\mathbf{c}_u = \{\mathbf{N}_1^{-1}\mathbf{M}_1^{-1}\mathbf{p}',\ \mathbf{N}_2^{-1}\mathbf{M}_1^{-1}\mathbf{p}',\ \mathbf{N}_3^{-1}\mathbf{M}_2^{-1}\mathbf{p}'',\ \mathbf{N}_4^{-1}\mathbf{M}_2^{-1}\mathbf{p}''\}$$

where $\mathbf{c}_u$ is a column vector of size $4v \times 1$. Note that, these previous steps can be can be executed offline, i.e, before the end of the billing interval.
- Step 4: The utility sends the encrypted price vector $\mathbf{c}_u$ to the GW.

3) Gateway: The GW can generate the electricity bill for the owner of SM_i as follows. During the billing interval, the GW stores all the ciphertexts received from SM_i, i.e., the GW stores $\{\mathbf{c}_{it} : 1 \leq t \leq T-1\}$. At the last reporting period T of the billing interval, the GW receives $\mathbf{c}_{iT}$ from SM_i. The GW computes $\mathbf{c}_i$ which is the encryption of the total consumption reported by SM_i during the billing interval as follows.

$$\begin{aligned} \mathbf{c}_i &= \sum_{t=1}^{T} \mathbf{c}_{it} \\ &= \begin{Bmatrix} \sum_{t=1}^{T} \mathbf{w}'_{it}\mathbf{A}_i\mathbf{N}_1 & , & \sum_{t=1}^{T} \mathbf{w}'_{it}\mathbf{B}_i\mathbf{N}_2 & , \\ \sum_{t=1}^{T} \mathbf{w}''_{it}\mathbf{C}_i\mathbf{N}_3 & , & \sum_{t=1}^{T} \mathbf{w}''_{it}\mathbf{D}_i\mathbf{N}_4 & , \end{Bmatrix} \end{aligned}$$

After receiving the encrypted price vector $\mathbf{c}_u$ from the utility, the GW can compute the electricity bill of SM_i using a single product operation between $\mathbf{c}_i$ and $\mathbf{c}_u$. The correctness proof is as follows.

$$
\begin{aligned}
\text{Bill}_i &= \mathbf{c}_i\mathbf{c}_u \\
&= \sum_{t=1}^{T} \mathbf{w}'_{it}\mathbf{A}_i\mathbf{M}_1^{-1}\mathbf{p}' + \sum_{t=1}^{T} \mathbf{w}'_{it}\mathbf{B}_i\mathbf{M}_1^{-1}\mathbf{p}' \\
&+ \sum_{t=1}^{T} \mathbf{w}''_{it}\mathbf{C}_i\mathbf{M}_2^{-1}\mathbf{p}'' + \sum_{t=1}^{T} \mathbf{w}''_{it}\mathbf{D}_i\mathbf{M}_2^{-1}\mathbf{p}'' \\
&= \sum_{t=1}^{T} \mathbf{w}'_{it}\left(\mathbf{A}_i + \mathbf{B}_i\right)\mathbf{M}_1^{-1}\mathbf{p}' \\
&+ \sum_{t=1}^{T} \mathbf{w}''_{it}\left(\mathbf{C}_i + \mathbf{D}_i\right)\mathbf{M}_2^{-1}\mathbf{p}'' \\
&= \sum_{t=1}^{T} \mathbf{w}'_{it}\mathbf{M}_1\mathbf{M}_1^{-1}\mathbf{p}' + \sum_{t=1}^{T} \mathbf{w}''_{it}\mathbf{M}_2\mathbf{M}_2^{-1}\mathbf{p}'' \\
&= \sum_{t=1}^{T} \mathbf{w}'_{it}\mathbf{p}' + \sum_{t=1}^{T} \mathbf{w}''_{it}\mathbf{p}'' \\
&= \mathbf{w}_i\mathbf{p}
\end{aligned}
$$

Note that the vector $\mathbf{w}_i = \mathbf{r}_i + \mathbf{m}_i = \sum_{t=1}^{T} \mathbf{r}_{it} + \sum_{i=1}^{T} \mathbf{m}_{it}$. However, since the masks are generated such that $\sum_{t=1}^{T} \mathbf{m}_{it} = 0$, then $\mathbf{w}_i = \mathbf{r}_i = \sum_{t=1}^{T} \mathbf{r}_{it}$. Therefore, the product of $\mathbf{w}_i$ and $\mathbf{p}$ represents the product of the price vector $\mathbf{p}$ by $\mathbf{r}_i$ which is the total power consumption of SM_i during the billing interval, i.e, the power consumption of each dimension is multiplied by the electricity price of that dimension. The basic billing case can be easily achieved by setting the same price for all the power consumption dimensions while creating the price vector $\mathbf{p}$. Therefore, MDMS can achieve the functional requirement (F2).

VI. Privacy Preservation Analysis

In this section, we investigate the security and privacy preservation of our scheme.

A. Privacy preservation of consumers' power consumption data against individual attackers

As discussed in subsection V-C, the reading vector $\mathbf{r}_{it}$, that should not be accessed by any entity in the network, is masked using the masking vector $\mathbf{m}_{it}$ to produce the masked vector $\mathbf{w}_{it} = \mathbf{r}_{it} + \mathbf{m}_{it}$ that is encrypted to produce the ciphertext $\mathbf{c}_{it} = \{\mathbf{w}'_{it}\mathbf{A}_i\mathbf{N}_1 \ , \ \mathbf{w}'_{it}\mathbf{B}_i\mathbf{N}_2 \ , \ \mathbf{w}''_{it}\mathbf{C}_i\mathbf{N}_3 \ , \ \mathbf{w}''_{it}\mathbf{D}_i\mathbf{N}_4\}$. The vector $\mathbf{s}$ is used as splitting indicator to split $\mathbf{w}_{it}$ into $\mathbf{w}'_{it}, \mathbf{w}''_{it}$. The secret key $\mathcal{K}_i = \{\mathbf{A}_i\mathbf{N}_1, \ \mathbf{B}_i\mathbf{N}_2, \ \mathbf{C}_i\mathbf{N}_3, \ \mathbf{D}_i\mathbf{N}_4\}$ is used to encrypt $\mathbf{w}'_{it}, \mathbf{w}''_{it}$. Our scheme is build on kNN encryption scheme whose security has been formally proven in the known ciphertext model [25]. Thus, without the knowledge of the master key $\mathcal{MK} = \{\mathbf{M}_1, \mathbf{M}_2, \ \mathbf{N}_1, \ \mathbf{N}_2, \ \mathbf{N}_3, \ \mathbf{N}_4\}$, the content of ciphertext cannot be recovered. Therefore, privacy preservation of $\mathbf{w}_{it}$, and in turns $\mathbf{r}_{it}$, can be achieved against both, the GW that receives the individual ciphertexts and the external adversary $\mathscr{A}$ who can eavesdrop on the communication to obtain any individual ciphertext.

In addition, meters do not use a shared key for the encryption process as done in [24]. Instead, each meter receives a unique encryption key from the KDC generated from the master key set $\mathcal{MK}$ as in [26]. Thus, a meter SM_ℓ that has an encryption key $\mathcal{K}_\ell = \{\mathbf{A}_\ell\mathbf{N}_1, \ \mathbf{B}_\ell\mathbf{N}_2, \ \mathbf{C}_\ell\mathbf{N}_3, \ \mathbf{D}_\ell\mathbf{N}_4\}$ cannot decrypt the ciphertext $\mathbf{c}_{it}$ generated by another meter SM_i [26]. Therefore, the ciphertext of SM_i cannot be decrypted by other meters in the network.

Although the utility has the master key set $\mathcal{MK}$ that can decrypt any individual or aggregated ciphertext, it has no access to the individual ciphertext $\mathbf{c}_{it}$. Also, it has no access to all the masking vectors used by all meters. Therefore, MDMS can protect the individual power consumption data of the consumers against individual attackers, i.e., MDMS satisfies the security/privacy requirement (SP1).

B. Privacy preservation of consumers' power consumption data against colluding attackers

Different from singular attacks in which a single attacker tries to reveal the power consumption data of any consumer, we consider in this subsection collusion attacks in which the adversary can collude with other entities in the AMI networks. Assume that an adversary $\mathscr{A}$ can control the utility, the GW, and a set of meters of size $|\mathbb{SM}_{\mathscr{A}}|$. In order to reveal the power consumption data of a meter, first $\mathscr{A}$ obtains from the GW the ciphertext $\mathbf{c}_{it}$ of the SM_i. Then, $\mathscr{A}$ uses the master key $\mathcal{MK}$ of the utility to decrypt $\mathbf{c}_{it}$ and recover the masked vector $\mathbf{w}_{it} = \mathbf{r}_{it} + \mathbf{m}_{it}$.

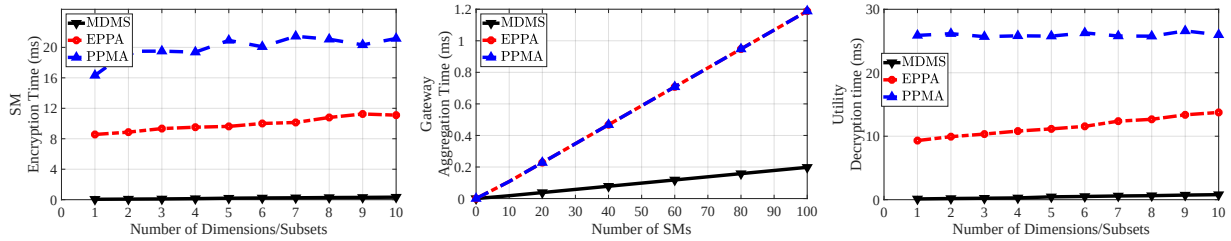
In order to recover the reading vector $\mathbf{r}_{it}$, $\mathscr{A}$ removes the masking vector $\mathbf{m}_{it}$. As mentioned in step 2 in subsection V-C, SM_i shares a pairwise secret masks with all other SMs in $\mathbb{SM}$ in an efficient and offline way. Therefore, as long as $|\mathbb{SM}_{\mathscr{A}}| < (|\mathbb{SM}| - 1)$, $\mathscr{A}$ cannot completely remove all the masks added to $\mathbf{m}_{it}$, and thus the reading vector $\mathbf{r}_{it}$ is still masked by the remaining $(|\mathbb{SM}| - 1) - |\mathbb{SM}_{\mathscr{A}}|$ masks.

In [30], a formal security proof is given to prove that the masked power consumption data is protected as long as the mask size is chosen properly, and masks are generated using a pseudorandom function. Therefore, MDMS can resist collusion attacks in which $\mathscr{A}$ controls the utility, GW and a set of $(\mathbb{SM} - 2)$ meters and thus MDMS can satisfy the security/privacy requirement (SP3).

The colluding set $\mathbb{SM}_{\mathscr{A}}$ can contain up to $(|\mathbb{SM}| - 2)$. It cannot contain all the $(|\mathbb{SM}| - 1)$ meters, i.e, all but one are colluding, since in this case $\mathscr{A}$ can obtain the aggregated power consumption data from the utility and simply recover the power consumption data of the honest meter by subtracting the consumption of the other colluding $(|\mathbb{SM}| - 1)$ meters from the aggregation result.

C. Confidentiality of the aggregated power consumption data and subsets' sizes

The aggregated vector, $\mathbf{r}_{at}$, that contains all the aggregated power consumption data as well as all the subsets' sizes should not be accessed by any entity in the system except the utility. As mentioned in subsection V-D, the GW receives all the individual ciphertexts and computes the aggregated ciphertext $\mathbf{c}_{at}$. Also, the external adversary $\mathscr{A}$ can eavesdrop on all the exchanged messages to obtain a copy of all the individual



(a) Smart meter computations. (b) Gateway computations. (c) Utility Computations.

Fig. 4. Computation overhead comparison

ciphertexts and thus $\mathscr{A}$ can compute the aggregated ciphertext $\mathbf{c}_{at}$ as well. Since the aggregation of all the ciphertexts results in removing all the masking vectors added by all meters, $\mathbf{c}_{at}$ becomes the encryption of the aggregated vector $\mathbf{r}_{at}$. Therefore, the GW and $\mathscr{A}$ can obtain the ciphertext $\mathbf{c}_{at}$ of the vector $\mathbf{r}_{at}$ that should be protected from any unauthorized access. However, $\mathbf{c}_{at}$ is still a ciphertext encrypted under the master key set $\mathcal{MK}$, which has been shown to be secured as discussed in subsection VI-A and the meters' keys cannot decrypt it. Thus, neither the GW nor $\mathscr{A}$ can learn any information about the aggregated power consumption data of each subset nor the subsets' sizes. Therefore, MDMS can satisfy the security/privacy requirement (SP2).

VII. Performance Evaluation

In this section, we evaluate the performance of MDMS in terms of the computation overhead required by each entity and the communication overhead incurred between each two entities in the network. We compare the performance of MDMS with EPPA [12] and PPMA [14].

A. Computation Overhead

The computation overhead is defined as the processing time required by each node in the network. These nodes are smart meters, the GW and the utility. In MDMS, we use the efficient ID-based aggregate signature scheme proposed in [21] which requires a fixed number of bilinear pairing operations regardless of the number of the individual signatures to verify the aggregate signature, i.e., the signature verification complexity is $\mathcal{O}(1)$. On contrary, EPPA [12] uses the signature scheme proposed in [31] which requires a number of pairing operations that increases linearly with the number of individual signatures to verify the aggregate signature, i.e., the signature verification complexity is $\mathcal{O}(n)$. Therefore, for the fairness of comparison, we do not consider the overhead of signing and verifying messages in MDMS and other schemes.

To evaluate the computation overhead, we implemented MDMS, EPPA, and PPMA, using Python charm cryptographic library [32] running on an Intel Core i7-4765T 2.00 GHz and 8 GB RAM. We used super-singular elliptic curve (SS512 curve) with the symmetric Type 1 pairing to realize the bilinear pairing operation. The size of the parameter q is 512 bits. All the experiments' results are as follows.

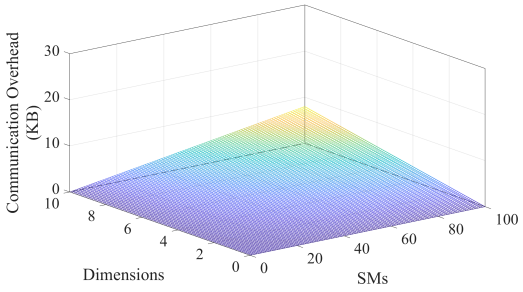
Figure 4a gives the computation overhead for each meter versus the number of data dimensions/subsets to be encrypted. As shown in the figure, the computation overhead required by each meter in MDMS increases slightly as the number of data dimensions/subsets increase. For instance, it increases from 52 μs when reporting one dimensions/subsets to 330 μs when reporting 10 dimensions/subsets. This is because the size of the vector to be encrypted $\mathbf{w}_{it}$ increases linearly with the number of dimensions/subsets, and thus more arithmetic addition and multiplication operations are needed during the vector encryption process. For EPPA [12], as the number of dimensions increases, EPPA requires more time to build the super-increasing sequence and then encrypt the sequence using Paillier cryptosystem. Therefore, the computation overhead of each meter in [12] also increases linearly with the number of data dimensions to be reported. On contrary, the computation overhead of each meter in PPMA [14] is almost constant. This is because the encryption process in PPMA requires only two exponentiation operations over $\mathbb{Z}_{n^2}$ regardless of the number of subsets represented in the message. The figure shows that MDMS reduces the computation overhead by more than 90% as compared to [12], [14]. This is because the encryption process in MDMS requires only efficient arithmetic addition and multiplication operations compared to the Paillier encryption time required in [12], [14]. Therefore, MDMS is more suitable than other schemes for the resource-constrained meters due to its lower computation overhead.

Figure 4b gives the computation overhead of the GW to aggregate ciphertexts versus the number of meters. As shown in the figure, EPPA and PPMA have exactly the same aggregation time because the GW aggregates Paillier ciphertexts in the two schemes. The figure also shows that MDMS is six times faster than other schemes because the aggregation process in MDMS requires only efficient addition operations.

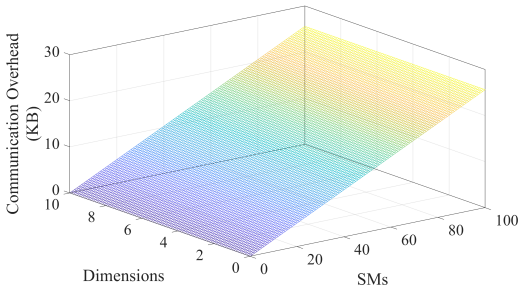
Figure 4c gives the computation overhead required by the utility versus the number of dimensions/subsets represented in the plaintext. The figure shows that MDMS is the most efficient compared to the schemes in [12], [14]. This is because in MDMS only one vector decryption operation, that includes arithmetic addition and multiplication, is required which is much more efficient than Paillier decryption operation used in [12], [14]. To sum up, MDMS outperforms other schemes in terms of computation overhead on each entity in the network.

B. Communication Overhead

The communication overhead is measured by the size of transmitted messages between the network entities. In specific, the communication overheads to be measured are for the messages sent from smart meters to GW and from GW to the

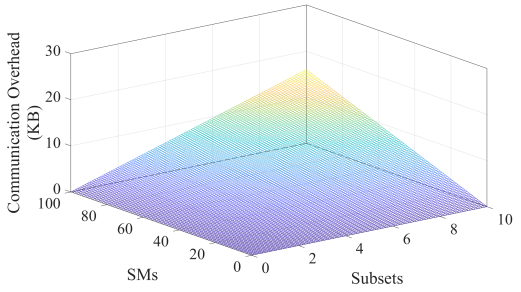


(a) MDMS.

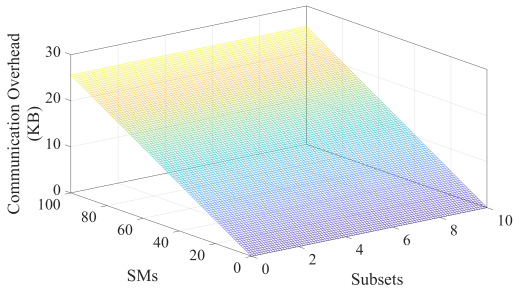


(b) EPPA [12].

Fig. 5. Communication overhead comparison for multi-dimensional data collection.



(a) MDMS.



(b) PPMA [14].

Fig. 6. Communication overhead comparison for multi-subset data collection.

utility. However, in all schemes under consideration, the GW aggregates all the ciphertexts to a single ciphertext. Therefore, in our comparison, we will focus on comparing the ciphertext size of each scheme.

In MDMS, the ciphertext is a vector of size $4v = 8mn$ elements. Consider that each element in the ciphertext is represented with 2 bytes, therefore, the ciphertext size in MDMS becomes $16mn$ bytes. For EPPA [12] and PPMA [14], the ciphertext size is the same as in Paillier cryptosystem which is 512 bytes (2,048 bits). However, EPPA only allows multi-dimensional data collection whereas PPMA allows only multi-subset data collection. For fair comparisons, we consider the following cases.

- **Case 1.** We compare MDMS to the EPPA scheme in [12] for multi-dimensional data collection only.
- **Case 2.** We compare MDMS to the PPMA scheme in [14] for multi-subset data collection only.
- **Case 3.** We compare MDMS to EPPA and PPMA when used to collect multi-dimensional multi-subset data collection simultaneously. For PPMA, to simultaneously support multi-dimensional and multi-subset data collection, an additional paillier ciphertext is sent for every data dimension. Similarly, for EPPA, an additional paillier ciphertext is sent for every additional subset.

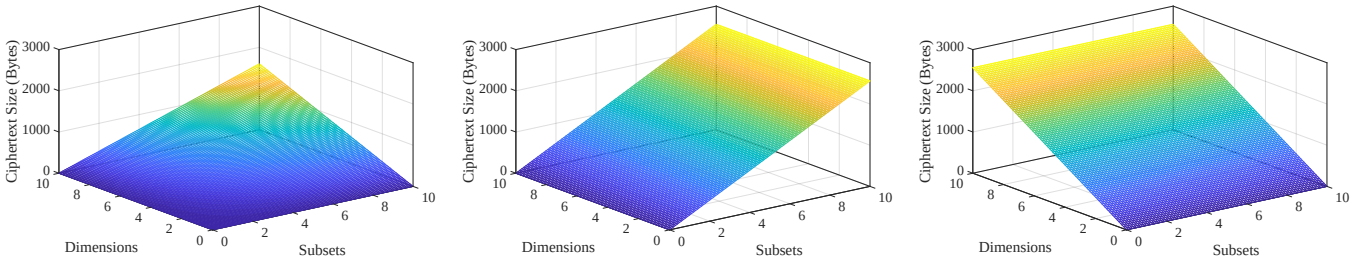
Case 1: Multi-dimensional data collection. In this case, the ciphertext size of MDMS can be reduced to $8m$ instead of $16mn$. There is a reduction factor of $(2n)$ as there is no need to report n elements for subsets readings nor n elements for the corresponding subsets' indicators in this case. We plot the SM-to-GW communication overhead in terms of the number of meters and the number of data dimensions. Fig. 5a gives the communication overhead of MDMS, whereas Fig. 5b gives the communication overhead of EPPA.

As shown in Fig. 5a, the total communication overhead in MDMS increases with both the number of meters and the data dimensions, whereas Fig. 5b shows that the total communication overhead in EPPA increases only with the number of meters. However, the figures show that MDMS always has a lower communication overhead compared to EPPA. This is because the single ciphertext size in MDMS is extremely low compared to that in EPPA.

Case 2: Multi-subset data collection. In this case, the ciphertext size of MDMS can be reduced to $16n$ instead of $16mn$. There is a reduction factor of (m) as there is no need to report m multi-dimensional data in this case. We plot the SM-to-GW communication overhead in terms of the number of meters and the number of subsets. Fig. 6a gives the communication overhead of MDMS, whereas Fig. 6b gives the communication overhead of PPMA.

As shown in Fig. 6a, the total communication overhead in MDMS increases with both the number of meters and subsets, whereas Fig. 6b shows that the total communication overhead in PPMA increases only as the number of meters inreases. However, the figures show that MDMS always has lower communication overhead compared to PPMA for the same reason as case 1.

Case 3: Simultaneous Multi-dimensional Multi-subset data collection. In this case, the ciphertext size of MDMS is $16n$. We plot the size of a single ciphertext sent by a smart meter versus the number of data dimensions and the number of subsets in Fig. 7a for MDMS, Fig. 7b for EPPA, and Fig. 7c for PPMA. The figures show that MDMS has the lowest ciphertext size when compared to EPPA and PPMA which in turns results in an improved communication overhead. As shown in the figure, for 10 data dimensions with 10 subsets to be reported within each dimension, MDMS can achieve a reduction of 40% in the ciphertext size.



(a) MDMS. (b) EPPA [12]. (c) PPMA [14].

Fig. 7. Ciphertext size comparison for all schemes when used for multi-dimensional multi-subset data collection.

VIII. Conclusion

In this paper, we have proposed MDMS, an efficient and privacy-preserving data collection scheme for AMI networks that allows multi-dimensional and multi-subset data collection. Based on, ID-based cryptography, the kNN encryption technique, and data masking technique, we have developed MDMS to achieve the functional and privacy requirements for the multi-dimensional multi-subset data collection in AMI networks. Also, MDMS was designed to all the utility to delegate electricity bill computation to local gateways to consider the network scalability in terms of the number of consumers. Our privacy analysis have demonstrated that MDMS can preserve the privacy of the consumers against individual and collusion attackers because no one can access the power consumption data of the individual consumers. Moreover, our performance evaluations demonstrated that MDMS is more computationally and bandwidth-wise efficient compared to relevant schemes in the literature. For the resource limited smart meters, MDMS is 8 times faster than closely similar existing schemes. For the communication overhead, MDMS can reduce the communication overhead by approximately 40%.

IX. Acknowledgements

This work was supported by the US National Science Foundation under Grants 1619250. The first author would like to thank the support of the University Research Council, University of Central Arkansas. The statements made herein are solely the responsibility of the authors.

References

[1] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 60–65, April 2011.

[2] U.S. Departmrnt of Energy, "The smart grid," https://www.smartgrid.gov/the_smart_grid/smart_grid.html, [Online; accessed 31-May-2019].

[3] Constance Douris, "Balancing Smart Grid Data and Consumer Privacy," http://www.lexingtoninstitute.org/wp-content/uploads/2017/07/Lexington_Smart_Grid_Data_Privacy-2017.pdf, 2017, [Online; accessed 31-May-2019].

[4] A. Paverd, A. Martin, and I. Brown, "Security and privacy in smart grid demand response systems," *Proceedings of the International Workshop on Smart Grid Security*, pp. 1–15, 2014.

[5] X. Li, X. Liang, R. Lu, X. Shen, X. Lin, and H. Zhu, "Securing smart grid: cyber attacks, countermeasures, and challenges," *IEEE Communications Magazine*, vol. 50, no. 8, 2012.

[6] K. Zhang, R. Lu, X. Liang, J. Qiao, and X. S. Shen, "PARK: A privacy-preserving aggregation scheme with adaptive key management for smart grid," *Proceedings of the IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 236–241, Aug 2013.

[7] M. Nabil, M. Ismail, M. M. Mahmoud, W. Alasmary, and E. Serpedin, "PPETD: Privacy-preserving electricity theft detection scheme with load monitoring and billing for ami networks," *IEEE Access*, vol. 7, pp. 96 334–96 348, 2019.

[8] A. Alsharif, M. Nabil, S. Tonyali, H. Mohammed, M. Mahmoud, and K. Akkaya, "EPIC: Efficient privacy-preserving scheme with EtoE data integrity and authenticity for AMI networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3309–3321, April 2019.

[9] 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.

[10] J. Ni, K. Alharbi, X. Lin, and X. Shen, "Security-enhanced data aggregation against malicious gateways in smart grid," *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2015.

[11] A. Alsharif, M. Nabil, M. Mahmoud, and M. Abdallah, "Privacy-preserving collection of power consumption data for enhanced AMI networks," *Proceedings of the 25th International Conference on Telecommunications (ICT)*, pp. 196–201, June 2018.

[12] R. Lu and X. Liang and X. Li and X. Lin and X. Shen, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, Sept 2012.

[13] R. Lu, K. Alharbi, X. Lin, and C. Huang, "A novel privacy-preserving set aggregation scheme for smart grid communications," *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 2015.

[14] S. Li, K. Xue, Q. Yang, and P. Hong, "PPMA: Privacy-preserving multisubset data aggregation in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 462–471, 2018.

[15] U.S. Environmental Protection Agency, "Electricity customers," https://www.epa.gov/energy/electricity-customers, 2018, [Online; accessed 31-May-2019].

[16] E. Oriero and M. A. Rahman, "Privacy preserving fine-grained data distribution aggregation for smart grid AMI networks," *Proceedings of the IEEE Military Communications Conference (MILCOM)*, 2018.

[17] H. Mohammed, S. R. Hasan, and M. A. Rahman, "Load control and privacy-preserving scheme for data collection in ami networks," *arXiv preprint arXiv:1807.11565*, 2018.

[18] C. Chen, S. Kishore, and L. V. Snyder, "An innovative RTP-based residential power scheduling scheme for smart grids," *Proceddings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, pp. 5956–5959, 2011.

[19] A. Sherif, A. Alsharif, M. Mahmoud, M. Abdallah, and M. Song, "Efficient privacy-preserving aggregation scheme for data sets," *Proceedings of the 25th International Conference on Telecommunications (ICT)*, pp. 191–195, June 2018.

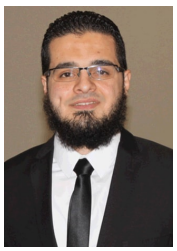
[20] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in *Theory of Cryptography Conference*. Springer, 2005, pp. 325–341.

[21] C. Gentry and Z. Ramzan, "Identity-based aggregate signatures," *Proceedings of the International workshop on public key cryptography*, pp. 257–273, 2006.

[22] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," *Proceedings of the Annual international cryptology conference*, pp. 213–229, 2001.

[23] M. Nabil, A. Alsharif, A. Sherif, M. Mahmoud, and M. Younis, "Efficient multi-keyword ranked search over encrypted data for multi-data-owner settings," *Proceedings of the IEEE International Conference on Communications (ICC)*, May 2018.

[24] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222–233, Jan 2014.
[25] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure kNN computation on encrypted databases," *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, pp. 139–152, 2009.
[26] H. Li, D. Liu, Y. Dai, and T. H. Luan, "Engineering searchable encryption of mobile cloud networks: when QoE meets QoP," *IEEE Wireless Communications*, vol. 22, no. 4, pp. 74–80, August 2015.
[27] A. Alsharif, M. Nabil, M. M. E. A. Mahmoud, and M. Abdallah, "EPDA: Efficient and privacy-preserving data collection and access control scheme for multi-recipient ami networks," *IEEE Access*, vol. 7, pp. 27 829–27 845, 2019.
[28] A. Sherif, A. Alsharif, J. Moran, and M. Mahmoud, "Privacy-preserving ride sharing organization scheme for autonomous vehicles in large cities," *Proceedings of the IEEE 86th Vehicular Technology Conference, IEEE VTC2017-Fall*, September 2017.
[29] A. Sherif, A. Alsharif, M. Mahmoud, and J. Moran, "Privacy-preserving autonomous cab service management scheme," *Proceedings of the 3rd Africa and Middle East Conference on Software Engineering*, December 2017.
[30] A. Unterweger, S. Taheri-Boshrooyeh, G. Eibl, F. Knirsch, A. Küpçü, and D. Engel, "Understanding game-based privacy proofs for energy consumption aggregation protocols," *IEEE Transactions on Smart Grid*, 2018.
[31] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and verifiably encrypted signatures from bilinear maps," *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 416–432, 2003.
[32] J. A. Akinyele, C. Garman, I. Miers, M. W. Pagano, M. Rushanan, M. Green, and A. D. Rubin, "Charm: a framework for rapidly prototyping cryptosystems," *Journal of Cryptographic Engineering*, vol. 3, no. 2, pp. 111–128, 2013.



Ahmad Alsharif (M'18) received the Ph.D. in Electrical and Computer Engineering from Tennessee Tech. University in May 2019. He received the B.Sc. and M.Sc. degrees with honors in Electrical Engineering from Benha University, Egypt in 2009 and 2015, respectively. In 2009, he was one of the recipients of the young innovator award from the Egyptian Industrial Modernisation Centre. Currently, Dr. Alsharif is an assistant professor at the University of Central Arkansas. His research interests include security and privacy in smart grid, vehicular Ad Hoc networks, multihop cellular wireless networks.

Mahmoud Nabil is an Assistant Professor in the department of Electrical and Computer Engineering, North Carolina A and T University, USA. He received his Ph.D. degree in Electrical and Computer Engineering from Tennessee Tech University, Cookeville, Tennessee, United States in August 2019. He received his B.S. degree and M.S. degree with honors in Computer Engineering from Cairo University, Egypt in 2012 and 2016, respectively. Mahmoud research interests include security and privacy in smart grid, machine learning applications, vehicular Ad Hoc networks, and blockchain applications.

Ahmed Sherif (M'19) is an Assistant Professor in the School of Computing Sciences and Computer Engineering at the University of Southern Mississippi (USM). He received his Ph.D. degree in Electrical and Computer Engineering from Tennessee Tech University, Cookeville, Tennessee, USA in August 2017. He received his M.Sc. degree in Computer Science and Engineering from Egypt-Japan University of Science and Technology (E-JUST) in 2014. His research interests include security and privacy-preserving schemes in Autonomous Vehicles (AVs), Vehicular Ad hoc Networks (VANETs), and Smart Grid network.

Mohamed Mahmoud received PhD degree from the University of Waterloo in April 2011. From May 2011 to May 2012, he worked as a postdoctoral fellow in the Broadband Communications Research group - University of Waterloo. From August 2012 to July 2013, he worked as a visiting scholar in University of Waterloo, and a postdoctoral fellow in Ryerson University. Currently, Dr Mahmoud is an associate professor in Department Electrical and Computer Engineering, Tennessee Tech University, USA. His research interests include security and privacy preserving schemes for smart grid communication, mobile ad hoc, sensor, and delay-tolerant networks. Dr. Mahmoud has received NSERC-PDF award. He won the Best Paper Award from IEEE International Conference on Communications (ICC'09), Dresden, Germany, 2009.

Min Song (SM'10, F'18) joined Stevens Institute of Technology in July 2018 as Professor and Chair of the Department of Electrical and Computer Engineering. Before joining Stevens, he was the David House Professor, Chair of the Computer Science Department and Professor of Electrical and Computer Engineering at Michigan Tech from 2014 to 2018. He was also the founding director of the Michigan Tech Institute of Computing and Cybersystems. Prior to joining Michigan Tech, Min served as a program director with the National Science Foundation (NSF) from 2010 to 2014. Mins professional career comprises 28 years in academia, government, and industry. Throughout his career, Min has published more than 165 technical papers and held various leadership positions. He served as TPC Co-Chair for many IEEE conferences including ICC and GLOBECOM. He has been serving as a member of the IEEE INFOCOM Steering Committee. He is the recipient of NSF CAREER Award in 2007 and NSF Directors Award in 2012. Min is an IEEE Fellow.