Towards Efficient Privacy-Preserving Data Aggregation for Advanced Metering Infrastructure

Navid Alamatsaz[1], Arash Boustani[2], Nima Alamatsaz[3], Ashkan Boustani[4]

[1,2]Department of Electrcial Engineering and Computer Science, Wichita State University, Wichita, KS, USA.

[3] Department of Biomedical Engineering, New Jersey Institute of Technology, Newark, NJ, USA.

[4] Department of Statistics and Mathematics, University of Red Crescent Society of Iran, Mashad, Iran.

Abstract

Recent changes to the existing power grid are expected to influence the way energy is pro-vided and consumed by customers. Advanced Metering Infrastructure (AMI) is a tool to incorporate these changes for modernizing the electricity grid. Growing energy needs are forcing government agencies and utility companies to move towards AMI systems as part of larger smart grid initiatives. The smart grid promises to enable a more reliable, sustainable, and efficient power grid by taking advantage of information and communication technologies. However, this information-based power grid can reveal sensitive private information from the user's perspective due to its ability to gather highly-granular power consumption data. This has resulted in limited consumer acceptance and proliferation of the smart grid. Hence, it is crucial to design a mechanism to prevent the leakage of such sensitive consumer usage information in smart grid. Among different solutions for preserving consumer privacy in Smart Grid Networks (SGN), private data aggregation techniques have received a tremendous focus from security researchers. Existing privacy-preserving aggregation mechanisms in SGNs utilize cryptographic techniques, specifically homomorphic properties of public-key cryptosystems. Such homomorphic approaches are bandwidth-intensive (due to large output blocks they generate), and in most cases, are computationally complex. In this paper, we present a novel and efficient CDMA-based approach to achieve privacy-preserving aggregation in SGNs by utilizing random perturbation of power consumption data and with limited use of traditional cryptography. We evaluate and validate the efficiency and performance of our proposed privacy-preserving data ag-gregation scheme through extensive statistical analyses and simulations.

Keywords

Smart Grid; Data-oriented Privacy; Secure data Aggregation; Spread Spectrum.

1 Introduction

A series of power surges over a twelve-second period triggered a cascade of shutdowns in the US and Ontario on August 14, 2003. The result was the biggest blackout in North

DOI: 10.5121/ijcnc.2017.9508

American history. 61800 megawatts of power were lost to over 50 million people. Studies showed that the outage was because of lack of real-time monitoring and diagnosis and failure in proper load balancing [2]. Recently, *Smart Grid* has been proposed as the next generation power grid. A Smart Grid is an electrical grid that leverages communication technologies and information processing to gather, process, and act on collected information to improve reliability, efficiency, economics, and sustainability of the power grid in generation, transmission, and distribution [3]. This information-based power grid will help the *Utility Companies (UC)* to act on consumer information gathered from *Smart Meters (SM)* at the user's premises. The two-way communication capability will enable functions such as demand-response, demand-dispatch, self-monitoring, and self-diagnosis for the existing power grid [4]. It also promises reduced prices through dynamic pricing schemes, wide penetration of renewable resources such as wind and solar, and fewer power outages [5]. The topic of smart grid has attracted researchers to study various aspects of modernizing the electricity grid. The research community has been studying miscellaneous subjects such as communication technologies and infrastructure [3, 6, 7, 8, 9], legal and policy concerns [10, 11], reliability, failure diagnosis and recovery [12, 13, 14], demand-response, demand-dispatch, load shaping, and peak-shaving [15, 16, 17], data aggregation [3, 18, 19, 20, 21, 22, 23] and, last but not the least, security and privacy [4, 5, 3, 24, 25, 26].

Advanced Metering Infrastructure (AMI) are systems that measure, gather, analyze energy usage, and communicate with metering devices such as water meters, gas meters, heat meters, and electricity meters. This communication is either on request or on a predetermined schedule. Government agencies and utilities are adopting AMI systems as part of the deployment of the smart grid. AMI improves current *Advanced Meter Reading (AMR)* technology by enabling two-way communications between the meter and the utility. This allows UCs to send commands to the meters for different purposes, such as time-of-use pricing information, demand-response actions, or remote disconnects [8].

Although AMI provides the UC with state-of-the-art capabilities, having access to fine-grained consumer usage data can reveal information regarding the private lives of its users. For instance, it can be easily determined if a residential house is vacant or not by observing the fine-grained energy consumption patterns [27]. It is also possible to track the location of the residents of a household based on the appliance they are using [28]. Insurance companies can monitor and track eating, sleeping, and possibly exercise habits of a household [29, 30]. In 2009, the Dutch Parliament prohibited the utilization of smart meters because of privacy issues. It is worth mentioning that in *Smart Grid Networks (SGN)*, data-oriented privacy is more of interest, as opposed to context-oriented privacy, because it deals with private consumer data. There are also many cyber security related challenges for the deployment of the Smart Grid [3]. This "Internet-like distributed power grid" is vulnerable to many known and unknown cyber security attacks [31]. The security threats to the Smart Grid can target the confidentiality and the integrity of the gathered fine-grained user data. They can also threaten the availability of the power grid. Computerworld [32] reports more than 170 outages caused by cyber-security attacks. It should go without saying that without appropriate security and privacy-preserving techniques, large-scale deployment and consumer-acceptance of the Smart Grid paradigm is difficult.

In general, data aggregation techniques are utilized to significantly reduce the volume of traffic being transmitted in an SGN by compressing data in the intermediate nodes (also called aggregators). Aggregation is an important technique for preserving network re-

sources, such as bandwidth and energy [33]. Also, it is deployed as a common approach to preserve data privacy against external adversaries as the aggregation process compresses large inputs to small outputs at the intermediate aggregators. However, this can lead to several new vulnerabilities against potential internal adversaries, such as the aggregator node itself. Thus, it is of paramount importance to design appropriate mechanisms for privacy-preserving data aggregation [54]. Earlier privacy-preserving approaches have primarily used cryptographic techniques such as homomorphic encryption and secure multiparty computation in order to preserve user privacy while aggregating usage data [35]. These approaches, although providing strong guarantees of confidentiality, are very heavy from a computational and communicational stand-point and may not be feasible on low-end smart meters with limited computation capabilities [62]. Considering the huge scale of future smart meter deployment and the granularity of the data being gathered, existing communication networks will have difficulty handling this data because of resource constraints such as network capacity (bandwidth) [66, 67, 68]. Homomorphic cryptosystems usually generate an output of a huge fixed-length as compared with the data generated by smart meters. This ciphertext can be up to one hundred times larger than the actual smart meter data [3]. Given the frequency of the data being sent and possible bandwidth scarcity, this can lead to unacceptable delay and network overhead [66].

In this paper, we investigate the feasibility of existing privacy-preserving data aggregation approaches. We devise a novel, efficient, and feasible (from a communications perspective) data aggregation mechanism for SMs using coding theory, *spread spectrum communications (SSC)*, and *random perturbation* techniques [36, 37]. We also evaluate the privacy protection level of our proposed scheme with well-established information-theoretic and statistical tools [38, 64, 39, 40]. Finally, we validate the performance of our aggregation mechanism by means of simulations.

The rest of the paper is organized as follows. Related work in the literature and background on existing secure aggregation schemes is outlined in Section 2. The network and adversary model assumed in this work along with basics of SSC are presented in Section 3. Our proposed perturbation-based privacy-preserving aggregation utilizing SSC is outlined in Section 4. Evaluation and simulation results are discussed in Section 5. We conclude the paper with a summary of contributions and results in Section 6.

2 Background and Related Work

In this section, we outline mechanisms in the literature for privacy-preserving data aggregation in SGNs and also study some data aggregation methods in other networking infrastructure with similar constraints such as *Wireless Sensor Networks (WSN)*.

2.1 Homomorphic Encryption for Data Aggregation

A public-key cryptosystem is known to have homomorphic properties if $E(m_1 \diamond m_2) = E(m_1) \triangle E(m_2)$, where E is the encryption function, $\diamond$ and $\triangle$ are two mathematical operations, and m_1, m_2 are two input messages. In other words, a homomorphic property enables certain mathematical operations on the plaintext by performing specific operations on the ciphertext without observing any intermediate results in plaintext. Based on the supported operations, homomorphic cryptosystems fall into two broad categories: par-

tially homomorphic and fully homomorphic. Partially homomorphic cryptosystems only support either addition or multiplication, or in some cases polynomials up to certain degrees, whereas fully homomorphic cryptosystems support both addition and multiplication [3, 26]. It goes without saying that fully homomorphic cryptosystems provide much more flexibility and have recently received significant attention [41, 42]. However, given their computational complexity, they are not widely used in practical applications yet. Well-known homomorphic cryptosystems include RSA [43], El Gamal [44], Paillier [42], Naccache-Stern [45], and Boneh-Goh-Nissim [46, 47].

In general, data aggregation techniques might support different aggregation functions such as sum, max, min, avg, median, and variance. However in SGNs, the UC is mostly interested in total consumption (sum) of a given neighborhood in a specific time period to enable functions such as demand-response, load-shaping, peak-shaving, and self-monitoring [4, 3, 15, 17]. Also, the average (avg) usage of each household might be of interest. Given that sum of consumed electricity of all smart meters in a residential neighborhood is required to be computed in a private fashion, the additive homomorphic property of the Paillier [42] cryptosystm can be useful. Also, the Boneh-Goh-Nissim cryptosystem [47, 26] (which is an extension of Paillier with bilinear groups) supports the additive homomorphic function. Rather than adding the consumption data in plaintext, one can multiply the encrypted values and then decrypt the result to get the addition of plaintext data. The Paillier encryption system works as explained in Protocol 1 (Key Generation), 2 (Encryption), and 3 (Decryption) [18]. As it can be observed, the sum of plaintext can be computed from multiplication of the ciphertext, i.e. $D(E(m_1).E(m_2)\ mod\ N^2) = (m_1 + m_2) mod\ N$ or $D(C_1.C_2\ mod\ N^2) = (m_1 + m_2) mod\ N$, where N is the modulus for encryption and decryption.

1 : **Generate** two large prime numbers p and q such that $gcd(p.q, (p-1), (q-1) = 1)$;
2 : **Calculate** $N = p.q$;
3 : **Calculate** $\lambda = lcm(p-1, q-1)$;
4 : **Select** a random number $g \in Z^*_{N^2}$;
5 : **if** (μ exists such that $\mu == (L(g^\lambda\ mod\ N^2))^{-1}\ mod\ N$ and $L(u) = \frac{u-1}{N}$ **then**
6 : (N, g) is the public key;
7 : (λ, μ) is the private key;
8 : **end if**
9 : **End**.

Protocol 1: Key Generation.

1 : Let $m \in Z_N$ be the plaintext;
2 : **Generate** random number $r \in Z^*_N$;
3 : **Calculate** ciphertext $c = (g^m.r^N)\ mod\ N^2$;
4 : **End**.

Protocol 2: Encryption.

1 : Let $c \in Z^*_{N^2}$ be the ciphertext;
2 : **Calculate** the plaintext $m = L(c^\lambda\ mod\ N^2).\mu\ mod\ N$;
9 : **End**.

Protocol 3: Decryption.

He et al. [26] present a secure data exchange scheme for the smart grid based on homomorphic properties of Goh cryptosystem [46]. Goh supports an arbitrary number of additions and a single multiplication on the ciphertext. It is worth noting that the aforementioned protocol is only a secure data communication scheme and does not address the problem of secure aggregation. Li et al. [18] utilize the homomorphic properties of Paillier to propose an incremental data aggregation scheme. In [18], every node passes its encrypted time-series data to its parent node on the aggregation tree. The parent node multiplies the received value into its own encrypted consumption data and passes the total result to the next parent node. Therefore, all the SMs participate in the aggregation without seeing any intermediate or final result. Garcia and Jacobs [48] present a privacy-preserving protocol using Paillier based on secret sharing. Their proposal hides consumption data from the UC as it receives random shares of data (instead of the entire data) which it cannot decrypt. The other nodes cannot retrieve meaningful information either since they only receive random shares. Kursawe et al. [49] propose two approaches to calculate total consumption in SGN. In their first approach, called *aggregation protocols*, smart metering data are masked in such a way that after summing the data from all smart meters masking values cancel each other out and the UC gets the total consumption information. In their second approach, named *comparison protocols*, they consider that the UC roughly knows the total consumption. Erkin and Tsudik [50] propose a cryptographic protocol based on a modified version of the Paillier cryptosystem to calculate the total consumption of all the SMs in a given neighborhood as well as a single SM in the AMI. Acs and Castelluccia [51] suggest a solution using masking and differential privacy and utilizing the homomorphic properties of a computationally-cheap cryptosystem for private data aggregation. Lu et al. [52] propose an *Efficient and Privacy-Preserving Aggregation (EPPA)* for smart grid communications by structuring multidimensional data and encrypting them with the Paillier cryptosystem. Erkin et al. [3] study different existing secure signal processing mechanisms in SGNs and compare different existing cryptographic methods in terms of computational complexity, efficiency, and imposed overhead.

It is worth noting that in WSNs another non-homomorphic, cryptographic approach has also been utilized; an intermediate node in the aggregation tree has to decrypt the data received from a downstream node, then aggregate the data according to the aggregation function, for instance sum, and finally encrypt the output of the aggregation function before forwarding the result to the up-stream node on the tree. Such schemes have several shortcomings, the most important of which is that they do not protect the privacy of the transmitted data from the neighboring sensor nodes. All neighbors share pairwise keys and are able to decrypt the incoming data. Hence, if the neighboring sensor node is honest-but-curious or if it is compromised and monitored by the adversary, the data in transit can be easily intercepted.

2.2 Non-homomorphic Private Data Aggregation

A common path to privacy-preserving aggregation in WSNs is perturbing the raw data being transmitted by introducing a random noise [36, 37, 54, 61]. He et al. [54] propose two approaches to privacy-preserving data aggregation in WSNs. The basic idea of their first approach, *Cluster-based Data Aggregation (CPDA)*, is to introduce noise to the raw data sensed by the sensor node, such that this noise will be cancelled out in the aggregation operation resulting in an accurate aggregate value. The main idea of their second proposed method, *Slice-Mix-AggRegaTe (SMART)*, is to slice original data into pieces and recombine

them randomly. Next, the authors further improve their protocol to *iPDA* which preserves the integrity of the data on top of its privacy [34]. In another perturbation-based effort, Zhang et al. [61] propose *Generic Privacy Preservation Solutions(GP^2S)* for approximate aggregation. In their proposed technique, the values of the data transmitted in a WSN are generalized such that individual data content cannot be decrypted. However, the aggregator can still calculate an estimate of the data distribution, and hence, approximately compute the aggregate value. Zanjani et al. [55, 56] propose a new energy-efficient aggregation mechanism for WSNs using the concepts of coding theory. The sensor nodes are assigned unique *Orthogonal Chip Sequences (OCS)* that are used to code and send their data on the CDMA channel. The authors claim that, by utilizing *ESTOC*, data integrity can be protected while aggregating. Also, *ESTOC* reduces *Bit Error Rate (BER)* and interference caused by simultaneous transmission of nodes. Yan et al. [19] propose a secure in-network data aggregation scheme to aggregate the data from smart appliances inside a *Home Area Network (HAN)* utilizing the properties of SSC for efficient aggregation. The authors only utilize OCSs for data aggregation and not for providing any security guarantees. They use Message Authentication Codes (MAC) for checking the authenticity of the transmitted data. However, confidentiality and integrity of the data is not protected. In our work, we propose a secure aggregation scheme based on the properties of OCSs to preserve the confidentiality of the transmitted data without relying on traditional cryptographic techniques.

2.3 Discussion

In the homomorphic encryption-based approaches discussed in [3, 18, 26, 48, 49, 50], we observe that the power-usage information is generally of small size (e.g. 20 bits) [4, 52]. However, the plaintext input size of most existing homomorphic cryptosystems is huge [3, 52], for example 2048 bits for the widely-used Paillier cryptosystem [42, 48, 50, 52]. As a result, the input data has to be padded before encryption and the size of the output is also large. Given the high frequency of data collection and the number of deployed smart meters, this will result in unacceptable communication overhead on the network, and also high processing burden on the smart meters with limited computational capabilities [52, 62]. Aggregation schemes that construct and utilize the spanning-tree, for instance by Li et al. [18], also do not consider performance issues. The processing and communication overhead makes the protocol less suitable in practical implementations. Moreover, depending on the depth of the spanning tree of the network, there can be large delays between the time power consumption data is reported by the meters and the time the aggregated data is received at the UC. In approaches proposed in [34, 54], the perturbed or the sliced data need to be encrypted before being sent to the neighbors. However, the key-distribution for such symmetric pair-wise encryption is non-trivial. In other words, any two node in the network will share symmetric keys which will result in a key distribution complexity of order $O(n^2)$, where n is the number of nodes in the network. Moreover, this encryption can put extra burden on the nodes with limited capabilities. Phulpin et al. [57] study the efficiency and benefits of network coding in both Power Line Communications (PLC) and wireless SGNs. The authors also show that using coding theory in SGN reduces the delay by decreasing the number of time slots and saves energy by reducing the number of transmissions.

Based on the aforementioned observations, designing an efficient privacy-preserving technique for aggregating SM data without using traditional crypto primitives with homomorphic properties seems to be necessary. We are proposing a privacy-preserving ag-

gregation scheme using coding theory, spread spectrum communications, and statistical perturbation in order to efficiently aggregate power usage while improving network performance and decreasing unnecessary communication and computation loads on the SGN. Our contention-free scheme will also decrease the delay, BER, and interference. Our contributions are twofold: First, we introduce a simple, yet efficient, approach to perturb user data before aggregation in order to preserve user privacy. Second, we propose a secure aggregation scheme, *AgSec*, using SSC. Finally, we assess the privacy level and the performance of our scheme through analytical evaluations and simulations.

3 Network Architecture

3.1 Network and Communication Model

Communication standards and technology to be used in the future smart grid and AMI is an ongoing debate. There are various communication options proposed for the smart grid including fiber optics, copper-wire line, power line communications, and miscellaneous wireless technologies. We consider the widely used wireless architecture for the deployment of SGN [8]. The wireless communication between SMs, which are organized into groups called *clusters*, and the aggregator or *Cluster Head (CH)* uses IEEE 802.15.4 or Zigbee due to characteristics such as low power, short delay, self-organization, scalability, and high security [8]. The aggregated data will be forwarded from the CH to the UC using a dedicated point-to-point link.

Figure 1 depicts the assumed three-level hierarchical network architecture. The communication between the UC and the i^{th} aggregator (CH) is denoted as UA_i. Similarly $AS_{i,j}$ represents the communication between the i^{th} aggregator and the j^{th} smart meter in the i^{th} cluster. Also there exists a separate out-of-band *control and signaling* channel between the i^{th} aggregator and the j^{th} smart meter in the i^{th} cluster referred to as $CC_{i,j}$. The signaling and control messages, which are used in the initialization phase, are discussed in detail in Section 4.1. The Zigbee medium access protocol on all *AS* channels is CDMA. Also, all UA communications are on a dedicated point-to-point channel. Our signaling channel uses a low-range wireless technology such as *IEEE 802.15.4* or *IEEE 802.11*. The main advantage of Wi-Fi over Zigbee is its high data rate. However, Wi-Fi's high energy consumption is an issue that should be considered. The Zigbee and Wi-Fi alliances have been working towards designing a standard that promotes Zigbee to work on Wi-Fi, called *Smart Energy 2.0* [8]. Finally, the i^{th} aggregator uses a CDMA broadcast channel BC_i to distribute the perturbation information. n OCSs are used to broadcast random noise information on BC_i. These random numbers will be utilized by SMs to perturb their time-series data.These n random numbers are placed in a $[\quad]_{i\times j=n}$ *Perturbation Matrix*, where n is the number of SMs in the cluster. Every element of this matrix is coded with a unique OCS as described in Section 4.2. Figure 3 illustrates the components implemented in different network entities.

3.2 Communications on the CDMA Channel

All communications take place over four separate channels, as discussed in Section 3.1. All smart meter data from the smart meter to the aggregator are sent over the CDMA-based data channel, represented as the *AS* channel (in Fig. 1). The OCSs for encoding data transmission on the *AS* channel are generated using the Golay or PCC code generation

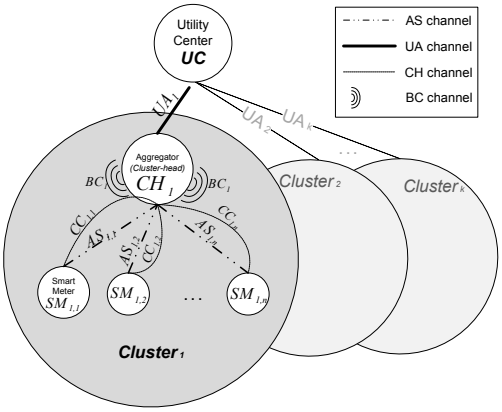

Figure 1: Network Architecture.

algorithms [58, 59]. These OCSs will be used to spread the data as explained later in Section 4.4. Golay OCSs can be generated recursively, as shown in Eqn. 1.

$$C_L = \begin{bmatrix} C_{\frac{L}{2}} & \bar{C}_{\frac{L}{2}} \\ C_{\frac{L}{2}} & -\bar{C}_{\frac{L}{2}} \end{bmatrix}$$

$$C_L = [\, A_L \; B_L \,], \bar{C}_L = [\, A_L \; -B_L \,] \text{ and } C_1 = \bar{C}_1 = [-1] \tag{1}$$

In Eqn. 1, $L = 2^M$ is the total number of available OCSs (which is also equal to the OCS length), where $M \geq 1$ is the number of chips in each OCS. A_L and B_L are $L \times \frac{L}{2}$ sub-matrices. In recursive OCS generation algorithms such as Golay (or PCC), OCSs can be organized into groups called *flock* based on *chip pattern similarity* and *chip distance* between OCSs. In Fig. 2-a, we can see the different flocks for 16-chip OCSs. Both Golay and PCC algorithms are able to produce L OCSs with a length of L-chips. The PCC generator matrix is shown in Eqn. 2 and OCSs of 16-chip length generated using PCC are shown in Fig. 2-b. OCSs generated by PCC have a uniform distribution of 1's and -1's, in contrast to OCSs generated by Golay. This property, which will result in having equal number of 1's and -1's, makes data transmission using PCC more fault tolerant than Golay. We can use any OCS generator algorithm (synchronous or asynchronous) in our proposed method. However, PCC and Golay are preferred because of equality in OCS length and number of generated OCSs, and high level of orthogonality [59].

$$P_{4^n} = \begin{bmatrix} P_{4^{n-1}} & P_{4^{n-1}} & P_{4^{n-1}} & -P_{4^{n-1}} \\ -P_{4^{n-1}} & P_{4^{n-1}} & -P_{4^{n-1}} & -P_{4^{n-1}} \\ P_{4^{n-1}} & P_{4^{n-1}} & -P_{4^{n-1}} & P_{4^{n-1}} \\ P_{4^{n-1}} & -P_{4^{n-1}} & -P_{4^{n-1}} & -P_{4^{n-1}} \end{bmatrix}$$

$$\forall n \geq 1, P_1 = [-1] \tag{2}$$

Let us assume that time is divided into periods of random length denoted by a random variable ψ_τ. During each period, each smart meter is assigned a subset of OCSs for use in that period by the CH. The assignment happens over the *CC* signaling channel. The communications over the *CC* channels are secured, from possible sniffing nodes, using symmetric key cryptography and shared keys between SM and CH. The OCSs for each smart

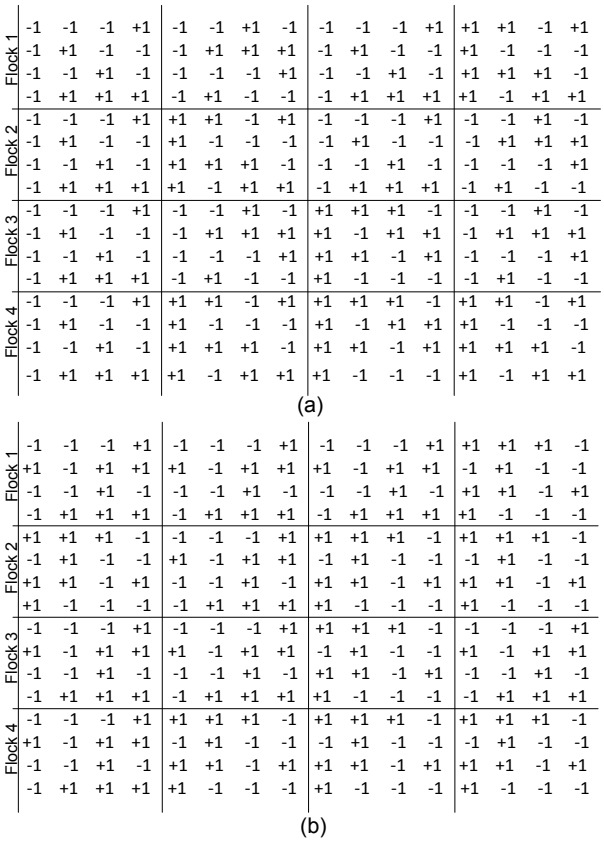

Figure 2: a) A 16-chip Golay OCS matrix. b) A 16-chip PCC OCS matrix.

meter are randomly selected by the CH from a large pool of available OCSs. Each smart meter will use the OCSs uniquely assigned to it in the time frame ψ_τ. In order to spread data bits on the *AS* data channel, the smart meter calculates the inner-product of every data-bit in appropriate OCS. Every single bit of data is coded independently with an OCS different from the previous and next data bit. This will build the foundation of our secure scheme as described in Section 4.4. It should be noted that it is possible for multiple smart meters to use the same OCS for data transmission in different parts of the network as long as their transmission ranges do not overlap and the SMs are in two diffrent clusters. This is required to make sure that the transmissions do not interfere with each other (in general, interference is anything that alters, modifies or disrupts a signal as it travels between a source and a receiver). The same CDMA concepts and principles are also deployed on the BC_i channel. This broadcast channel is used by the CH to advertise perturbation data to the SMs, as discussed in Section 4.2.

It should be noted that, before spreading the data on the CDMA channel using the introduced OCSs, a scrambling code is utilized between the sender and receiver for security purposes. This code, which is generally 2^{42} chips long , is referred to as the *Long Code*.

In order to appropriately use this long code, the sender and receiver must be synchronous with a GPS or *Coordinated Universal Time (UTC)* system [69].

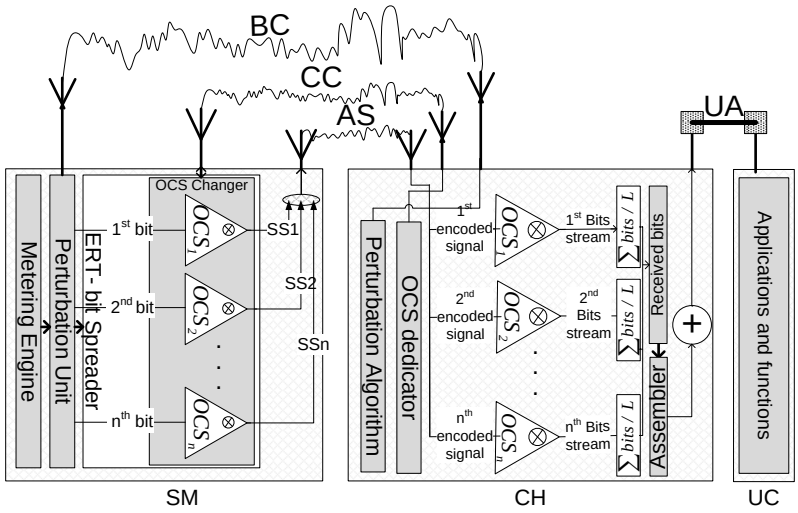

Figure 3: Entities used in the privacy-preserving aggregation.

3.3 Adversary Model

Based on their behavior, all entities in the proposed smart grid communication network can fall into one of the following three broad categories. (i) *honest* entities that fully follow the rules of the established protocol. (ii) *malicious* or *cheating* nodes that do not follow the protocol. Malicious behavior includes, but is not limitted to, insertion, deletion, and forging of messages in the system. (iii) *semi-honest* or *honest-but-curious* nodes that follow the defined protocols but they attempt to infer privacy-sensitive data from the input/output of the protocols and the intermediate data generated due to protocol execution. In our proposed scheme we consider the UC and the CH as honest-but-curious. In other words, they follow the established protocol but they can also try to infer privacy-sensitive information from the time-series data. The neighboring SMs are, generally, semi-honest. Our objective is to completely secure all the communications from malicious and semi-honest SMs and other adversarial nodes against possible sniffing, spoofing, and inference attacks and hence, maintain the consumers' privacy while still providing the UC with required aggregate values. Particularly, we are interested in protecting the system against the following attacks: (i) inference of individual data by CH and UC. (ii) eavesdropping (sniffing) by external adversaries. (iii) forging (spoofing) of smart meter data.

4 Privacy-Preserving Aggregation

4.1 Initialization Phase

Upon initial deployment, CH_i communicates control information to smart meter SM_j through $CC_{i,j}$. For each time duration ψ_τ, the CH assigns each smart meter, SM_j, a set

of attributes including, a temporary eight-bit identifier ($ID_{i,j}$) and a group of valid OCSs, denoted by $G^j_{\psi_\tau} = \{OCS^j_{1\psi_\tau}, OCS^j_{2\psi_\tau}, ..., OCS^j_{\zeta\psi_\tau}\}$. Also, the CH advertises the OCSs it is going to use for sharing perturbation information, denoted by $OCS_{(\lambda_1,\lambda_2,...,\lambda_n),\tau}$ for timeslot ψ_τ, on BC_i via the same $CC_{i,j}$, as will be discussed later in Section 4.2. These OCSs will be used by SMs to code/decode on the broadcast perturbation channel. The integrity, authenticity, and confidentiality of the communication between the CH and the SMs during the initialization phase are ensured using appropriate cryptographic techniques. In this phase, every smart meter gets the information required for data transmission on the CDMA channel and for data perturbation in the next t time-slots, as illustrated in Fig. 4. It should be noted that, as this is a one-time process in every t time slots and $\psi_t \gg \psi_\tau$, the imposed overhead is negligible. Also, we are not including any frame-level error checking mechanisms such as CRC because of the inherent fault-tolerance properties present in spread spectrum communications.

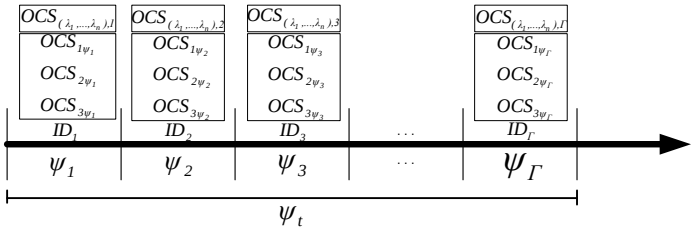

Figure 4: Initialization Parameters.

4.2 Privacy-Preserving via Random Noise Perturbation

Before discussing our secure aggregation protocol, we would like to introduce our random noise perturbation technique. Instead of aggregating the original smart meter data and sending the aggregate value to the UC, every smart meter utilizes a pseudo-random noise to perturb its data before aggregation. This perturbed data (instead of the original data) will be sent for aggregation to the CH. The received perturbed values P_i will be aggregated at the CH given the aggregation function in Section 4.4. Perturbation techniques in the literature usually follow two approaches. The basic idea of one group of such approaches is to add noise to the actual data such that the aggregator, or the CH in our case, can calculate an accurate aggregate value without inferring individual data transmitted by every node [54]. In a second similar direction, the data can be manipulated such that the aggregator can calculate an aggregate value which is an estimate of the histogram of data distribution rather than the actual aggregate value of the original data [61].

After all SMs are configured with appropriate OCS and ID information; they should start transmitting their readings periodically. Different time intervals for data reporting, ranging from 30 seconds to a few hours, could be found in the literature [4]. However, before transmitting, some noise should be added to this raw data. This random noise should be chosen in such a way that it does not affect the total aggregate value.

As noted earlier, in smart metering systems, the UC is generally interested in the output of two aggregation functions for a given neighborhood in a specific time period ψ_τ. First,

the sum of consumed electricity is desired, and second, the average consumption of every smart meter is of interest. These two values can help power companies plan accordingly for demand-response purposes. Based on these assumptions, our perturbation technique must be designed in such a way that the aggregator can calculate an accurate aggregate value while keeping individual meter readings confidential. Assume every $SM_{i,j}$ in cluster i has the data d_j to transmit. The sum and average of the data of all the SMs in this n smart meter cluster is:

$$SUM_i = d_1 + d_2 + ... + d_n = \sum_{j=1}^{n} d_j$$

$$AVG_i = \frac{d_1+d_2+...+d_n}{n} = \sum_{j=1}^{n} \frac{d_j}{n}$$

Now, assume that every SM adds a random value (noise) to its original data before transmission (How this noise is generated and distributed will be explained later in this section). We denote the perturbed data of SM_j by $P_j = d_j + \alpha_j$, where α_j is the random noise added to the raw data by SM_j. Hence, CH will be computing the sum of P_j's denoted by SUM_i':

$$\begin{aligned} SUM_i' &= (d_1 + \alpha_1) + (d_2 + \alpha_2) + ... + (d_n + \alpha_n) \\ &= \sum_{j=1}^{n} (d_j + \alpha_j) = \sum_{j=1}^{n} (p_j) \end{aligned}$$

In order for the CH to be able to calculate an accurate aggregate value we must have: $SUM_i = SUM_i'$ (and consequently $AVG_i = AVG_i'$). This implies that:

$$\sum_{j=1}^{n} \alpha_j = \alpha_1 + \alpha_2 + ... + \alpha_n = 0$$

Thus, for every given time period ψ_τ the CH must generate a series of random numbers that satisfy the above condition. These random numbers are advertised on the CDMA broadcast channel BC_i as an n element matrix where n is the number of SMs in cluster i. These n pseudo-random numbers are generated as follows. These constraints will guarantee that the summation of all the pseudo-random numbers is zero at all times.

1. If the number of SMs in the cluster is even (*n* is even), the CH will randomly generate $\frac{n}{2}$ positive integers α_j from the range $[0, max]$. Then, for every positive integer α_j it will place both α_j and $-\alpha_j$ in the perturbation matrix.

2. If the number of SMs in the cluster is odd (*n* is odd), the CH will randomly generate $\frac{n-3}{2}$ positive integers α_j from the range $[0, max]$. Then, for every positive integer α_j it will place both α_j and -α_j in the perturbation matrix. Next, it produces a positive random number α_σ and puts α_σ, $-\frac{\alpha_\sigma}{2}$, and $-\frac{\alpha_\sigma}{2}$ in the perturbation matrix (and hence having generated n random numbers).

After the perturbation matrix is generated by the CH, it should be advertised on BC_i. Every single element of this matrix, which includes a random number, will be encoded by an appropriate OCS (these OCSs are already shared in the initialization phase between the CH and SMs) and broadcast on the BC_i channel. $\xi(\alpha_j, OCS_{\lambda_j})$ denotes the j^{th} element of the matrix including pseudo-random number α_j encoded with OCS_{λ_j}. Every SM senses the channel, picks a random element of the matrix, decodes it with appropriate OCS (which it already learnt in the initialization phase) and uses that pseudo-random number to perturb its data. After the j^{th} element of the matrix is fetched and decoded, the SM will jam that element of the matrix (representing an invalid or already-used pseudo-random number) [60]. Assume SM_k has fetched and decoded pseudo-random number α_j spread with

OCS_{λ_j}. After this pseudo-random number α_j is used by SM_k, it needs to be jammed so that no other SM in the network uses the same α_j. In order for SM_k to generate the jamming signal, it transmits a packet with data value "all 1s" spread with OCS_{λ_j} (the same OCS that the pseudo-random number was encoded with), and with a higher transmit power. This will result in the corruption of α_j on the CDMA channel and will ensure that every α_j is used only by one smart meter, and hence, the summation of the added noise to the original data of all SMs in a given cluster is zero. It is worth mentioning that this jamming signal is transmitted without any transmitter-specific parameters, such as a source MAC address. This will ensure that the jamming signal cannot be linked to the transmitting SM, and thus, the pseudo-random numbers used by the smart meters are kept private and can be identified neither by the CH, nor by passive sniffing adversaries. To make the protocol more efficient, after α_j is replaced by all 1's, the CH can infer that this element of the matrix has been used, and hence, will stop advertising α_j. Consequently, SM_k will stop *jamming* on that specific OCS. Figure 5 illustrates the perturbation matrix.

As an alternative solution, after a smart meter fetches a pseudo-random number, it can send a packet on the control channel back to the CH indicating that pseudo-random number has been used. The sender of the packet has to be anonymized such that CH cannot distinguish which SM is using that pseudo random number. Different anonymization techniques (such as replacing the sender ID with a pseudonym) can be found in the literature [63]. In the anonymization process, the packets sent from SM to CH are anonymized, i.e., the user part (source) of each packet is replaced by a user pseudonym.

One-to-one Random Number Assignment: In order for the perturbation proposal to work as desired, we need to make sure that there is a one-to-one relationship between the random numbers α_i and the smart meters SM_j. This one-to-one assignment cannot be handled by the CH as it will result in compromising the privacy of SM data. Thus, it is crucial to design a mechanism to guarantee that every SM is using one unique random number and every random number is being used by atmost one SM. Let us assume that the SMs in a given cluster are time-synchronized. While the CH is advertising the random numbers matrix on BC_i channel, at the beginning of each time slot every SM accesses the data on each OCS with probability p and the SM will not read the data encoded with that specific OCS with probability $(1-p)$. A SM can use the accessed α only if no other SM has fetched the same α. Remember, after every α_i is fetched, the SM will send a jamming signal on that specific OCS; if more than two SMs are jamming the same OCS a collision is detected. This process is continued until all SMs have received one unique perturbation value. Suppose there are n' smart meters trying to access unique pesudo-random numbers at a given time instant. Then, the probability that accessing a given α is successful is the probability that only one of the SMs accesses that α and the other $(n'-1)$ SMs do not. The probability that an SM reads α is p; the probability that all other SMs do not read that α is $(1-p)^{(n'-1)}$. Therefore the probability that a given SM has a success is $p \times (1-p)^{(n'-1)}$. Because there are n' SMs, the probability that any one SM has a success is $n' \times p \times (1-p)^{(n'-1)}$.

4.3 Privacy Protection Evaluation

Many efforts in the past few years have been focused on designing privacy-preserving mechanisms for the smart grid. However, only a limited number of these works have presented an analytical framework to quantify the privacy leakage before and after the deploy-

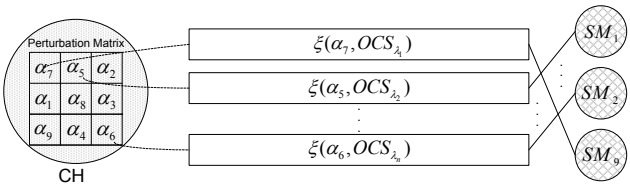

Figure 5: Perturbation Matrix.

ment of their privacy-preserving approach. Here, we introduce a simplistic certainty-based privacy analysis.

The notion of entropy by Shannon is a well-known measure of uncertainty in information theory [64]. The maximum uncertainty is achieved when entropy is maximized. Let X be a continuous random variable with probability density function (pdf) $f_X(x)$, then, the entropy of X is defined as follows:

$$H(X) = E[-\log f(X)] = \int_{-\infty}^{+\infty} [-\log f(x)] f(x) dx \tag{3}$$

It has been generally assumed that the electricity consumption patterns follow the Gaussian (normal) distribution [65]. Let $X \sim N(\mu, \sigma)$ denote the data generated by smart meters, where μ is the mean and σ is the standard deviation of the distribution. Then, the entropy of X is:

$$\begin{aligned} H(X) = E[-\log_e f(X)] &= \int_{-\infty}^{+\infty} [-\ln f_X(x)] f_X(x) dx \\ &= \int_{-\infty}^{+\infty} -\ln(\tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}) \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{+\infty} [\ln(\sqrt{2\pi}\sigma) + \tfrac{(x-\mu)^2}{2\sigma^2}] \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \ln(\sqrt{2\pi}\sigma) + \tfrac{1}{2\sigma^2} \int_{-\infty}^{+\infty} (x-\mu)^2 \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \ln(\sqrt{2\pi}\sigma) + \tfrac{1}{2\sigma^2}\sigma^2 \\ &= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \end{aligned} \tag{4}$$

Based on the above equation, the entropy of a Gaussian random variable only depends on the standard deviation σ and is independent of the mean μ. The data generated by the smart meters, based on our protocol, will be perturbed such that $\widehat{X} = X + \mathcal{A}$, where $\widehat{X}$ is a random variable denoting the perturbed data, and $\mathcal{A} \sim U(b, c)$ is a continuous random variable with a uniform distribution and the pdf $f_{\mathcal{A}}(\alpha) = \frac{1}{c-b}, \quad b \leq \alpha \leq c$ and $f_{\mathcal{A}}(\alpha) = 0$ otherwise, that models the generated perturbation data. The entropy of $\mathcal{A}$ is:

$$H(\mathcal{A}) = E[-\log_e f(\mathcal{A})] = \int_{-\infty}^{+\infty} [-\ln f_{\mathcal{A}}(\alpha)] f_{\mathcal{A}}(\alpha) d\alpha$$

$$= \int_{b}^{c} \frac{1}{c-b} \ln(c-b)\, d\alpha = \ln(c-b) \tag{5}$$

It should go without saying that increasing the range from which the random numbers are selected $(c-b)$ will increase entropy, and thus, decrease certainty of potential inference attacks. Now, we would like to see the result of this perturbation on the entropy of $\widehat{X}$. In general, assuming that X and $\mathcal{A}$ are two continuous random variables with the same range, we have [64]:

$$H(\widehat{X}) = H(X+\mathcal{A}) \geq max\{H(X), H(\mathcal{A})\}$$

$$\geq max\{\frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2), \ln(c-b)\} \tag{6}$$

As it can be concluded from the above equation, the entropy of $\widehat{X}$ is always greater than or equal to the maximum entropy of X and $\mathcal{A}$. Since the uniform distribution has the maximum entropy among all distributions, adding the smart meter data with uniformly distributed pseudo-random numbers will maximize entropy, minimize certainty, and hence, improve privacy. It should be noted that, here, we are not assuming any specific attack functions or adversarial strength. Given the a priori knowledge of the adversary, it might be able to infer information by observing $\widehat{X}$. In such a scenario, the entropy of the inferred information, denoted by the random variable Y, should be studied.

4.4 Proposed Secure Aggregation Protocol(AgSec)

After each SM adds appropriate noise to its original metering data, this perturbed data should be transmitted to the CH. In order to preserve data confidentiality against possible malicious entities and also other semi-honest smart meters and aggregators, we introduce a novel aggregation scheme that does not utilize cryptography and yet keeps the transmitted data secure. As discussed in Section 3.2, each node j is assigned a group of OCSs ($G^j_{\psi_\tau}$) for each time interval ψ_τ. The k^{th} bit of the (perturbed) data-stream generated by SM_j will be coded with $O^j_{(k \bmod g)}$, where g is the total number of OCSs assigned to SM_j in a given timeslot ψ_τ. The OCS $O_i(t)$ assigned to any SM_i at any instant of time t can be represented as shown in Eqn. 7.

$$O_i(t) = \sum_{j=0}^{L-1} O_{(j,i)}.p(t - jT_c) \tag{7}$$

In Eqn. 7, $p(t)$ is a rectangular pulse which is equal to 1 for $0 \leq t < T_c$ and zero otherwise. T_c is the chip duration of the OCS and $O_{(j,i)}$ is the j^{th} chip of the OCS assigned to SM_i (from the set of all OCSs C_L). The signal generated after encoding a data symbol of SM_i with the corresponding OCS is given by:

$$x_i(t) = d_i \sum_{j=0}^{L-1} O_{(j,i)}.p(t - jT_c) \qquad 0 \leq t < T_f \tag{8}$$

where, d_i is the data symbol of SM_i that needs to be encoded and $T_f = L.T_c$ is the duration of the encoded data symbol or data bit. The inner product of the sent bit with the OCS is

done bit-synchronously. Then, the overall transmitted signal $x(t)$ of all n SMs in a cluster can be given by Eqn. 9 [58].

$$x(t) = \sum_{i=0}^{n} x_i(t) \tag{9}$$

CH will receive a signal including all the bits transmitted by all the smart meters. The received signal will be decoded by CH using all valid OCSs that it initially assigned to the SMs. Since CH maintains a table of assigned OCSs (in the same order that was agreed in the initialization phase) and IDs to every single SM in the network, it is able to decode the data by using appropriate OCS for every bit. Hence, after decoding the received signal, CH has all individual (perturbed) data sent by all the SMs in the cluster. Then, it adds all the received data and sends the aggregate value to the UC on the point-to-point UA link. It should be noted again, the perturbation noise will be cancelled out upon addition. Our proposed secure aggregation technique is outlined in protocols 4, 5 and 6. (Even if data in transit could be decoded, it would still not be useful to the adversary as they are already perturbed.)

```
1 : Function (UA data transmission)
2 : While data on UA channels do
3 :     For all valid received aggregated data do
4 :         Collect all data values;
5 :     End For
6 : End While
7 : Utilize the aggregated data;
8 : End Function.
```

Protocol 4: UC function.

In protocol 4, the UC receives the aggregated data from the CH on the UA channel. Protocol 5 elaborates how CH generates and distributes OCSs (for aggregation and perturbation) to the SMs. Also, it shows how the data is despread, aggregated, and forwarded to the UC by CH. Finally, protocol 6 elaborates how SM receives the initialization information, perturbs data and transmits to the CH on the AS channel.

4.4.1 Security Analysis

Here, we would like to show that sniffing attacks against our CDMA-based aggregation are not feasible. This claim is based on the following considerations:

1. In any CDMA system, synchronous transmitters and receivers use a scrambling code, referred to as the *Long Code* or *Privacy Code*, which is used as a measure of security. This code is generally 2^{42} chips long and will return to its initial state after 41.43 days. For any sniffing adversary to decode the transmitted packets, it requires a prior knowledge of this long code [69].

2. Every P bits of data in the smart meter packet is encoded with sixty four possible OCSs resulting in L^P combinations (every SM packet is P bits long). Also, every one of these L^P combinations is a valid numeric value (assuming that smart grid data only contains numbers) that are indistinguishable from the adversary's perspective.

```
1 : Function (AS operation)
2 : For each each time perioud ψ_τ do
3 :    Generate the OCS table with Golay;
4 :    Function (Initialization);
5 :    For each each time period ψ_τ do
6 :        Generate the perturbation table and advertise on BC do;
7 :        For each advertised element on BC do;
8 :          If receive jamming signal on OCS_{λ_i} then;
9 :            Stop advertising on OCS_{λ_i};
10 :         End If
11 :       End For
12 :       Function(AS data transmission);
13 :    End For
14 :End For
15 :End Function
16 : Function (Initialization)
17 :Generate random IDs for SMs;
18 :Assign OCSs to each SN;
19 :End Function
20 : Function (AS data transmission)
21 : While data on AS channel do
22 :    For all valid OCSs do
23 :     Decode every received bit with appropriate OCS and reconstruct every SMs data;
24 :    End For
25 :    Calculate the SUM of all the received data;
26 :    Forward the aggregate value to the UC;
27 : End While
28 : End Function.
```

Protocol 5: CH function.

```
1 : While network is ON do
2 :    Function(BC data);
3 :    Function(Metering engine);
4 : End While
5 : End Function.
6 : Function (BC data)
7 : For OCS_{λ_j} do
8 :    Decode the received α_j on OCS_{λ_j};
9 :    Transmit a jamming signal on OCS_{λ_j} to jam α_j ;
10 : End While
11 : End Function.
12 : Function (Metering engine)
13 : While metering engine is ON do
14 :    Add α_j to the original data;
15 :    Encode the k^{th} of the perturbed data with O^j_{(k mod g)};
16 :    Spread the encoded data on the AS CDMA channel;
17 : End While
```

Protocol 6: SM function.

Now, assume that a packet is captured by a sniffer. Every bit of this packet will be spread with a 2^{42} bit long code and a L chip OCS. Given the length of the packet, this will result in $(2^{42} \times L)^P$ possible combinations which will be infeasible to decode using traditional brute force attacks. The only entity in the network that knows about the set of assigned OCSs to the smart meters is the CH. Hence, data confidentiality, to a great extent, will be preserved and privacy-sensitive information cannot be inferred by semi-honest and malicious entities.

5 Evaluation and Simulation Results

Below, we present a simple analysis that compares end-to-end and hop-by-hop delays in homomorphic approaches versus our proposed CDMA-based aggregation. We evaluate the performance of our aggregation scheme through extensive simulations.

5.1 Comparative Performance Evaluation by Numerical Analysis

As discussed in Section 2.1, existing secure aggregation schemes impose a significant communication and computation overhead on SGNs with limited capabilities. Private aggregation schemes based on the homomorphic properties of cryptosystems require fixed large size input blocks and are not ideally suited for small-sized data generated by SMs. The 20 to 30 bit [3] output data generated by SMs has to be padded, e.g., to 2048 bits for Paillier [42], before encryption. In our approach, by choosing OCSs with appropriate length, this overhead can be significantly reduced. Readers should note that in our scheme each bit will be spread to L bits after encoding.

In this section, we will numerically compare *End-to-End (ETE)* delay in our approach and homomorphic-based aggregation schemes. We are evaluating our results with clusters of ten and also twenty smart meters and assuming that each SM is assigned three OCSs to use in every given time slot. Given that each SM is assigned three OCSs, using an OCS with $L = 32$ and $L = 64$ will be ideal for each scenario, respectively. The OCS length L limits the maximum number of users per cluster to $\frac{L}{|G^j_{\psi_\tau}|}$. The total number of users in the network is independent of the OCS structure used. The transmission delay (D_T) for one SM can be calculated as:

$$D_T = \frac{(F + H_{ID}).L}{R} \tag{10}$$

where F is the frame length, H_{ID} is the ID header, L is the OCS length and R is the link bit-rate. Given Eqn. 10, the transmission delay using $L = 32$ and $L = 64$, assuming a 200 *kbps* ZigBee link, is 4.8 *ms* and 9.6 *ms*, respectively. However, using traditional homorphic cryptosystems as proposed by [18], the transmission delay(D_T) is:

$$D_T = \frac{(H_{ID} + D_C + T_{CRC})}{R} \tag{11}$$

where H_{ID} is the identifier header, D_C is the encrypted data (payload) and T_{CRC} is the error-checking trailer. Common SM and AMR systems generate data packets which contain a 24-bit meter ID (H_{ID}), a 22-bit meter reading and a 16-bit CRC checksum (T_{CRC}) [4]. This 22-bit meter reading is padded to 2048 bits before encryption and generates and output cipher of length 2048 bits (D_C). Based on these values, the transmission delay will be 10.44

ms for one SM. Another shortcoming of the privacy preserving homomorphic aggregation schemes, such as [18], is that every node's data should be passed hierarchically to the upper level node in the aggregation tree. This process continues until all the data is aggregated at the UC. However, this can increase the total delay which depends on the depth of the aggregation tree. Thus, if the depth of the aggregation tree is $\wp$, the total transmission delay will be $D_T \times \wp$. Given clusters of 10 or 20 SMs, in the worst case scenario, $\wp = 10$ and $\wp = 20$, and consequently $D_{T_\wp} = 104.4ms$ and $D_{T_\wp} = 208.8ms$, respectively. In the average case, the length of the aggregation tree, considering clusters of ten or twenty SMs, will be $\wp = 4$ and $\wp = 5$. Hence, transmission delay is $D_{T_\wp} = 41.76\ ms$ and $D_{T_\wp} = 52.2\ ms$, respectively. Our approach overcomes this issue as all nodes are able to transmit their data simultaneously and independently. This shows that our protocol is independent of the depth of the aggregation tree. Hence, using an OCS with appropriate length we are able to decrease the overhead significantly, as seen in Table 1. It should be noted that we are only considering the transmission delay. Moreover, given the high processing load and queuing delays due to the non-simultaneous transmission and high BER and retransmissions, the overall delay of the homomorphic approaches are too high compared with *AgSec*. Table 1 summarizes the transmission delay and total $communication\ overhead = \frac{Transmitted\ data}{Actual\ payload}$.

Table 1: Transmission Delay and Communication Overhead

	Agsec L=32 chips	**Agsec L=64 chips**	**Homomorphic (Paillier)**
D_T **for one SM** (ms)	4.8	9.6	10.44
D_T **for ten SM** (ms)	4.8	9.6	104.44
D_T **for twenty SM** (ms)	4.8	9.6	208.8
Communication Overhead	43.63	87.26	94.91

It is worth mentioning that Saputro and Akkaya [62] have analyzed the performance of homomorphic aggregation through extensive simulations. Not surprisingly, their results confirm our evaluation. The authors show that homomorphic encryption for data aggregation is very expensive in terms of communication overhead. They have also compared ETE homomorphic data aggregation with *Hop-by-Hop (HBH)* decrypt, aggregate, encrypt at intermediate aggregator nodes via regular stream-ciphers, such as RC-4. Surprisingly, both approaches show similar performance from a computation perspective (One multiplication in homomorphic ETE aggregation is as expensive as three operations in HBH aggregation: decrypt, add, encrypt) [62]. However, as our analysis also confirms, the authors show that ETE aggregation via homomorphic encryption generates extraordinarily large data which will result in unacceptable communication overhead on the SGN.

5.2 Simulation Results

We evaluate our proposed privacy-preserving aggregation protocol in a $100 \times 100\ km^2$ simulated metropolitan area with 50000 SMs. Our first goal is to verify the efficiency of our protocol in securely aggregating SM data as compared with existing approaches that employ homomorphic encryption for aggregation. One of the first observations we make is that, if appropriate parameters are chosen, our scheme performs more efficiently in terms of communication overhead and delay. Simulation parameters can be found in Table 2.

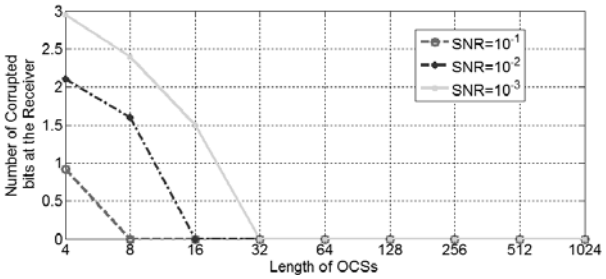

Figure 6: OCS Length versus Error.

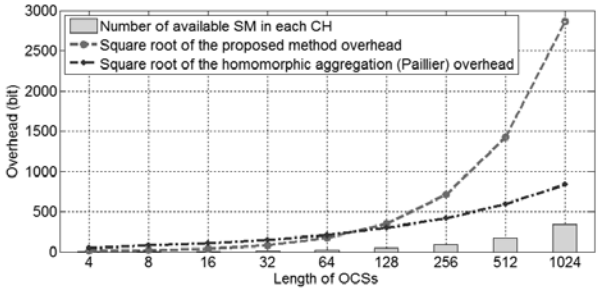

Figure 7: OCS Length versus Communication Overhead.

The 50000 SMs are clustered into groups of n SMs per cluster, where $n \leq \frac{L}{|G^j_{\psi_\tau}|}$ as every SM will be assigned ψ_τ OCSs out of the all L possible OCSs. One important aspect of the protocol that must be studied is the OCS length L which affects the number of SMs per cluster, tolerated error at the receiver, delay, and communication overhead. We observe that, at a constant SNR, the number of corrupted bits at the receiver decreases by increasing OCS length (Figure 6). As it can be clearly seen in Fig. 6, at $SNR = 10^{-3}$, if the OCS length is equal to or greater than 32 chips, there will be no error at the receiver. OCS lengths 16 and 8 will be ideal for $SNR = 10^{-2}$ and $SNR = 10^{-1}$, respectively. However, there is a trade-off between error and communication overhead. An increase in the OCS length will result in more communication overhead on the network. Our proposed scheme will outperform homomorphic aggregation, in terms of communication overhead, if the OCS length used is less than 128 chips. Figure 7 compares the communication overhead of our proposed CDMA-based aggregation with homomorphic aggregation schemes such as [18]. This confirms our analysis that an OCS length of 32 or 64 will be ideal in terms of error and communication overhead at $SNR = 10^{-3}$.

As mentioned earlier, the delay in ETE homomorphic encryption depends on the number of nodes and the depth of the aggregation tree. On the contrary, in our proposed scheme all the SMs are able to transmit their data independently and simultaneously. This will result in a considerable decrease in the end-to-end delay. Figure 8 compares the delays of our scheme with an ETE homomorphic approach such as [18]. As it can be clearly observed, our CDMA-based aggregation scheme significantly reduces delay.

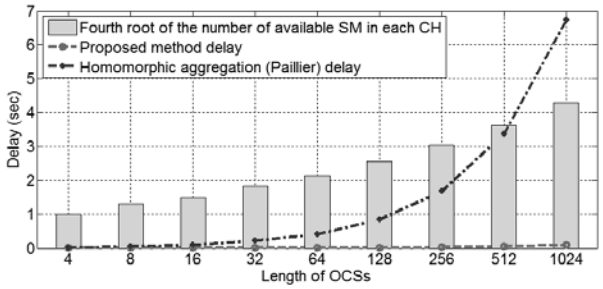

Figure 8: OCS Length versus Delay.

Table 2: Simulation Parameters

Parameter	**Value**
Network size	$100 \times 100\ km^2$
Cluster radius	100∼200 m
Number of SMs	50000
Number of SMs per cluster	$\lfloor \frac{L}{3} \rfloor$
AS Communication multiplexing	CDMA
OCS generator algorithm	4 to 1024 chips Golay OCSs
UA link	point-to-point
AS, BC, CC links	IEEE 802.15.4 Zigbee, FHSS, 2.4 to 2.48 GHz
AS Bit rate	200 Kbps
SM T_X and R_X power	100 mW, 20 dbm
Aggregator tree	fixed/static
CC security	public-key cryptography and digital signiture
Propagation model	free space

6 Conclusion

Existing approaches to privacy-preserving data aggregation in smart grid generally utilize the homomorphic properties of public-key cryptosystems. However, as we have thoroughly investigated, these approaches are expensive from a communication stand-point. In this paper, we proposed a two-step approach towards efficient private data aggregation in SGNs. First, we introduced a random perturbation technique which is used to statistically alter the time-series data of every SM such that individual consumption patterns could not be inferred and yet the sum and average values of the reported power consumption in a given neighborhood can be calculated accurately. Second, we proposed an efficient and secure data aggregation scheme which utilizes the properties of spread spectrum communications. Our evaluation and simulation results confirmed that our approach increases performance and decreases communication overhead on SGNs considerably, as compared with existing homomorphic aggregation schemes.

References

[1] Alamatsaz, N., Boustani, A., Jadliwala, M., Namboodiri, V. (2014) AgSec: Secure and Efficient CDMA-based Aggregation for Smart Metering Systems,Consumer Communications and Netwokring Conference (CCNC), 2014 11th IEEE, 102-108.

[2] IESO, Blackout 2003, url = http://www.ieso.ca/imoweb/EmergencyPrep/ blackout2003, 2012.

[3] Erkin, Z., Troncoso-Pastoriza, J.R., Lagendijk, R.L., Perez-Gonzalez, F. (2013) Privacy-preserving data aggregation in smart metering systems: an overview, Signal Processing Magazine, IEEE, volume 30, number 2, pages=75-86, ISSN=1053-5888

[4] Rouf, Ishtiaq and Mustafa, Hossen and Xu, Miao and Xu, Wenyuan and Miller, Rob and Gruteser, Marco, (2012) Neighborhood Watch: Security and Privacy Analysis of Automatic Meter Reading Systems, Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12, isbn = 978-1-4503-1651-4, Raleigh, North Carolina, USA,462–473, ACM, New York, NY, USA.

[5] Barenghi, A. and Pelosi, G. (2011) Security and Privacy in Smart Grid Infrastructures, Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on, 102-108, ISSN=1529-4188

[6] Van Engelen, A.G., Collins, J.S., (2010) Choices for Smart Grid Implementation, System Sciences (HICSS), 2010 43rd Hawaii International Conference on, ISSN=1530-1605.

[7] Bose, A., (2010) Smart Transmission Grid Applications and Their Supporting Infrastructure, Smart Grid, IEEE Transactions on, volume 1, number 1, pages 11-19, ISSN 1949-3053.

[8] Yu, F.R., Zhang, p., Weidong, X., Choudhury, P., (2011), Communication systems for grid integration of renewable energy resources, Network, IEEE, volume 25, number 5, pages 22-29, ISSN 0890-8044.

[9] Amin, R., Martin, J., Xuehai Z., (2012) Smart Grid communication using next generation heterogeneous wireless networks, Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on, pages 229-234.

[10] Tabors, R.D., Parker, G., Caramanis, M.C., (2010) Development of the Smart Grid: Missing Elements in the Policy Process, System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pages 1-7, ISSN 1530-1605.

[11] Schuler, R.E., (2010) Electricity Markets, Reliability and the Environment: Smartening-Up the Grid, System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pages 1-7, ISSN 1530-1605.

[12] Niyato, D., Wang, P., Hossain, E., (2012) Reliability analysis and redundancy design of smart grid wireless communications system for demand side management, Wireless Communications, IEEE, volume 19, number 3, pages 38-46, ISSN 1536-1284.

[13] Falahati, B., Yong F., Lei W., (2012) Reliability Assessment of Smart Grid Considering Direct Cyber-Power Interdependencies, Smart Grid, IEEE Transactions on, volume 3, number 3, pages 1515-1524, ISSN 1949-3053.

[14] Moslehi, K., Kumar, R., (2010) A Reliability Perspective of the Smart Grid, Smart Grid, IEEE Transactions on, volume 1, number 1, pages 57-64, ISSN 1949-3053.

[15] Shao, S., Pipattanasomporn, M., Rahman, S., (2011) Demand Response as a Load Shaping Tool in an Intelligent Grid With Electric Vehicles, Smart Grid, IEEE Transactions on, volume 2, number 4, pages 624-631, ISSN 1949-3053.

[16] Shao, S., Pipattanasomporn, M., Rahman, S., (2012) Grid Integration of Electric Vehicles and Demand Response With Customer Choice, Smart Grid, IEEE Transactions on,volume 3, number 1, pages 543-550, ISSN 1949-3053.

[17] Paschalidis, I.C., Binbin, L., Caramanis, M.C., (2012) Demand-Side Management for Regulation Service Provisioning Through Internal Pricing, Power Systems, IEEE Transactions on, volume

27, number 3, pages 1531-1539, ISSN 0885-8950.

[18] Li, F., Luo, B., Liu, P., (2010) Secure Information Aggregation for Smart Grids Using Homomorphic Encryption, Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, pages 327-332.

[19] Yan, Y., Qian, Y., Sharif, H., (2011) A Secure Data Aggregation and Dispatch Scheme for Home Area Networks in Smart Grid, Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE, pages 1-6, ISSN 1930-529X.

[20] Bartoli, A., Hernandez-Serrano, J., Soriano, M., Dohler, M., Kountouris, A., Barthel, D., (2010) Secure Lossless Aggregation for Smart Grid M2M Networks, Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, pages 333-338.

[21] Li, H., Lin, K., Li, K, (2011), Energy-efficient and High-accuracy Secure Data Aggregation in Wireless Sensor Networks, Elsevier Science Publishers B. V., volume 34, number 4, issn 0140-3664, pages 591–597.

[22] Weng, C., Li, M., Lu, X., (2008) Data Aggregation with Multiple Spanning Trees in Wireless Sensor Networks, Embedded Software and Systems, 2008. ICESS '08. International Conference on, pages 355-362.

[23] Mustafa, M., Zhang, N., Kalogridis, G., Fan, Z., (2014) DESA: A Decentralized, Efficient and SelectiveAggregation Scheme in AMI, Smart Grid, IEEE Transactions on.

[24] Fhom, H.S., Bayarou, K.M., (2011) Towards a Holistic Privacy Engineering Approach for Smart Grid Systems, Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on, pages 234-241.

[25] Line, M.B., Tondel, I.A., Jaatun, M.G., (2011) Cyber security challenges in Smart Grids, Innovative Smart Grid Technologies (ISGT Europe), 2011 2nd IEEE PES International Conference and Exhibition on, pages 1-8, ISSN 2165-4816.

[26] Xingze, H., Pun, M., Kuo, C.-C.J., (2012) Secure and efficient cryptosystem for smart grid using homomorphic encryption, Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES, pages 1-8.

[27] Lisovich, M. A., Wicker, S., (2008) Privacy concerns in upcoming residential and commercial demand-response systems, Clemson University Power Systems Conference.

[28] Lisovich, M.A., Mulligan, D.K., Wicker, S.B., (2010) Inferring Personal Information from Demand-Response Systems, Security Privacy, IEEE, volume 8, number 1, pages 11-20, ISSN 1540-7993.

[29] Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., Irwin, D., (2010) Private Memoirs of a Smart Meter, ACM BuildSys Work shop, isbn 978-1-4503-0458-0, Zurich, Switzerland, series BuildSys '10, pages 61–66, url= http://doi.acm.org/10.1145/1878431.1878446, New York, NY, USA.

[30] McDaniel, P., McLaughlin, S., (2009) Security and Privacy Challenges in the Smart Grid, Security Privacy, IEEE, volume 7, number 3, pages 75-77, ISSN 1540-7993.

[31] Cohen, F., (2010) The Smarter Grid, Security Privacy, IEEE, volume 8, number 1, pages 60-63, ISSN 1540-7993.

[32] Vijayan, j., (2010) Stuxnet renews power grid security concerns, Computerworld magazine.

[33] Li, N., Zhang, N., Das, S. K., Thuraisingham, B., (2009) Privacy Preservation in Wireless Sensor Networks: A State-of-the-art Survey, Elsevier Science Publishers B. V., Ad Hoc Network, volume 7, number 8, issn 1570-8705, pages 1501-1514.

[34] He, W., Nguyen, H., Liu, X., Nahrstedt, K., Abdelzaher, T., (2008) iPDA: An integrity-protecting private data aggregation scheme for wireless sensor networks, IEEE Military Communications Conference, pages 1-7.

[35] Lagendijk, R.L. and Erkin, Z. and Barni, M., (2013) Encrypted signal processing for privacy

protection: Conveying the utility of homomorphic encryption and multiparty computation, Signal Processing Magazine, IEEE, volume 30, number1, pages 82-105, ISSN 1053-5888.

[36] Agrawal, R., Srikan, R., (2000) Privacy-preserving data mining, SIGMOD, pages 49-54.

[37] Huang, Z., Du, W., Chen, B., (2005) Deriving Private Information from Randomized Data, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data SIGMOD '05, isbn 1-59593-060-4, pages 37-48, url = http://doi.acm.org/10.1145/1066157.1066163.

[38] Ross, S. M. (2009) Introduction to Probability and Statistics for Engineers and Scientists, Fourth Edition- Academic Press.

[39] Koralov, L. B., Sinai, Y. G., Theory of Probability and Random Processes,Second Edition-Springer.

[40] Cover, T., Thomas J., (2006) Elements of Information Theory, JohnWiley and Sons.

[41] Van Dijk, M., Gentry, C., Halevi, S., Vaikuntanathan, V., (2010) Fully Homomorphic Encryption over the Integers, Proceedings of the 29th Annual International Conference on Theory and Applications of Cryptographic Techniques EUROCRYPT'10, isbn 3-642-13189-1, 978-3-642-13189-9, French Riviera, France, pages 24-43, Springer-Verlag.

[42] Paillier, P., (1999) Public-key cryptosystems based on composite degree residuosity classes, EUROCRYPT.

[43] Rivest, R. L., Shamir, A., Adleman, L., (1978) A Method for Obtaining Digital Signatures and Public-key Cryptosystems, Commun. ACM, volume 21, number 2, issn 0001-0782, pages 120-126.

[44] El Gamal, T., (1985) A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms, Proceedings of CRYPTO 84 on Advances in Cryptology, isbn 0-387-15658-5, Santa Barbara, California, USA, pages 10-18, Springer-Verlag.

[45] Naccache, D., Stern, J., (1998) A New Public Key Cryptosystem Based on Higher Residues, ACM Conference on Computer and Communications Security, isbn 1-58113-007-4, pages 59-66, ACM.

[46] Goh, E.J., (2007) Encryption Schemes from Bilinear Maps, Department of Computer Science, Stanford University.

[47] Boneh, D.,Goh, E., Nissim, K., (2005) Evaluating 2-DNF Formulas on Ciphertexts, Proceedings of the Second International Conference on Theory of Cryptography, TCC'05, isbn 3-540-24573-1, 978-3-540-24573-5, pages 325–341, Springer-Verlag.

[48] Garcia F. D, Jacobs B., (2010) Privacy-friendly energy-metering via homomorphic encryption.

[49] Kursawe, K., Danezis, G., Kohlweiss, M., (2011) Privacy-friendly Aggregation for the Smart-grid, Proceedings of the 11th International Conference on Privacy Enhancing Technologies PETS'11, isbn 978-3-642-22262-7, Waterloo, ON, Canada, pages 175-191, url = http://dl.acm.org/citation.cfm?id=2032162.2032172, Springer-Verlag.

[50] Erkin, Z., Tsudik G., (2012) Private computation of spatial and temporal power consumption with smart meters ACNS.

[51] Cs G., Castelluccia C., (2011) I have a DREAM! (Differentially PrivatE smart Metering),ACM IH.

[52] Rongxing, Lu, Xiaohui, L., Xu, L., Xiaodong, L., Xuemin, S., (2012) EPPA: An Efficient and Privacy-Preserving Aggregation Scheme for Secure Smart Grid Communications, EEE Tran. on Parallel and Distributed Systems, volume 23, number 9, pages 1621-1631, ISSN=1045-9219.

[53] Plantard, T., Susilo, W., Zhang, Z., (2013) Fully Homomorphic Encryption Using Hidden Ideal Lattice, Information Forensics and Security, IEEE Transactions on, volume 8, pages 2127-2137, ISSN 1556-6013.

[54] Wenbo H., Xue L., Hoang N., Nahrstedt, K., Abdelzaher, T., (2007) PDA: Privacy-Preserving Data Aggregation in Wireless Sensor Networks, IEEE INFOCOM, pages 2045-2053, ISSN 0743-166X.

[55] Zanjani, M.B., Monsefi, R., Boustani, A., (2010) Energy efficient/highly secure data aggregation method using tree-structured orthogonal codes for Wireless Sensor Networks, Software Technology and Engineering (ICSTE), 2010 2nd International Conference on, volume 2, pages V2-260-V2-265.

[56] Zanjani, M.B., Boustani, A., (2011) Energy aware and highly secured data aggregation for grid-based asynchronous Wireless Sensor Networks, IEEE PacRim'11, pages 555-560, ISSN 1555-5798.

[57] Phulpin, Y., Barros, J., Lucani, D., (2011) Network coding in Smart Grids, Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on, pages=49-54.

[58] Chen, H. H., (2007) The Next Generation CDMA Technologies, John Wiley and Sons.

[59] Boustani, A., Sabet, J., Azizi, M., Mirmotahhary, N., Khorsandi, S., (2010) Persian Code: A new orthogonal spreading code generation algorithm for spread spectrum CDMA systems, Wireless Advanced (WiAD), 2010 6th Conference on, pages 1-5.

[60] Boustani, A., Alamatsaz, N., Jadliwala, M., Namboodiri, V. (2014) LocJam: A Novel Jamming-based Approach to Secure Localization in Wireless Networks, Consumer Communications and Netwokring Conference (CCNC), 2014 11th IEEE.

[61] Zhang, W., Wang, C., Feng, T., (2008) Generic Privacy-Preservation Solutions for Approximate Aggregation of Sensor Data (concise contribution), Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on, pages 179-184.

[62] Saputro, N., Akkaya, K., (2012) Performance Evaluation of Smart Grid Data Aggregation via Homomorphic Encryption, Wireless Communications and Networking Conference (WCNC), 2012 IEEE, pages 2945-2950, ISSN 1525-3511.

[63] Xu, D., Wang, Y., Shi, X., Yin, X., (2010) 802.11 User Anonymization, Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, pages 1-5.

[64] Gray, R. M., (2011) Entropy and Information Theory, Second Edition- Springer, isbn 978-1441979698.

[65] Sankar, L., Rajagopalan, S.R., Mohajer, S., Poor, H.V., (2013) Smart Meter Privacy: A Theoretical Framework, Smart Grid, IEEE Transactions on, volume 4, pages 837-846, ISSN 1949-3053.

[66] Karimi, B., Namboodiri, V., Jadliwala, M., (2013), On the scalable collection of metering data in smart grids through message concatenation, Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on, pp. 318-323, doi 10.1109/SmartGridComm.2013.6687977.

[67] Luan, W., Sharp, D., Lancashire, S., (2010) Smart grid communication network capacity planning for power utilities, in Transmission and Distribution Conference and Exposition, IEEE PES, pp. 1 4.

[68] Allalouf, M., Gershinsky, G., Lewin-Eytan, L., Naor, J., (2011) Data-qualityaware volume reduction in smart grid networks, in Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on, 2011, pp. 120125.

[69] Fazel, K., Kaiser, S., (2008) Multi-Carrier and Spread Spectrum Systems From OFDM and MC-CDMA to LTE and WiMAX, A John Wiley and Sons, Ltd, ISBN 978-0-470-99821-2.