Privacy-Preserving Data Sharing in Smart Grid Systems

Lei Yang, Hao Xue and Fengjun Li
EECS Department, The University of Kansas, Email: {lei.yang, haoxue, fli}@ku.edu

Abstract—**The smart grid systems aim to integrate conventional power grids with modern information communication technology. While intensive research efforts have been focused on ensuring data correctness in AMI data collection and protecting data confidentiality in smart grid communications, less effort has been devoted to privacy protection in smart grid data management and sharing. In smart grid data management, the Advanced Metering Infrastructure (AMI) collects high-frequency energy consumption data, which often contains rich inhabitant and lifestyle information about the end consumers. The data is often shared with various stakeholders, such as the generators, distributors and marketers. However, the utility may not have consent of the users to share potentially sensitive data. In this paper, we develop comprehensive mechanisms to enable privacy-preserving smart data management. First, we analyze the privacy threats and consumer identifiability issues associated with high-frequency AMI data. We then present the first solution based on data sanitization, which eliminates sensitive/identifiable information before sharing usage data with external peers. Meanwhile, we present solutions based on secure multi-party computing to enable external peers to perform aggregate/statistical operations on original metering data in a privacy-preserving manner. Experiments on real-world consumption data demonstrate the validity and effectiveness of the proposed solutions.**

I. Introduction

Envisioned as the next-generation power grid, the smart grid (SG) modernizes the existing power grid with bidirectional communication and pervasive computing capabilities for smart generation, distribution, management and consumption. In smart grid systems, high-frequency measurement data (e.g., register readings and time-interval consumption data), power quality data (e.g., voltage or current phase angle) and event data (e.g., outage alert) are collected from millions of smart meters and sent to the meter data repository at utilities. These data are expected to play a key role in supporting intelligent management applications (e.g., load forecasting, demand management, outage management, energy theft detection, etc.) and improving smart grid stability and energy efficiency. However, when the power grid evolves to become "smart", new security challenges have emerged. The security concerns come from the enlarged attack surface, for instance, smart meters are placed in physically insecure locations that are easily accessible to the adversaries, and thus are subject to attacks from physically tampering with meter reading (e.g., meter invasion) to manipulating measurement data in communication channel between meter processor and the embedded sensor. Moreover, the upgraded connectivity makes smart meters susceptible to cyber attacks in which adversaries may compromise smart meters or eavesdrop the communication.

While intensive research efforts have been focused on data correctness and trustworthiness in AMI data collection, less effort has been devoted to data privacy protection. In SG, energy consumption data that contains rich information about end consumers is collected at a much higher frequency than before. Without proper protection, realtime fine-grained metering data may disclose sensitive information about the consumers and expose them to a variety of privacy threats. For example, information about the lifestyle of the inhabitants can be inferred from high-resolution metering data via non-intrusive appliance load monitoring (NIALM). In [1], the power consumption data is correlated to appliance usage to associate power events with automated appliances activities and inhabitant's activities. Such privacy-sensitive household data may be used by third-party industries to profile energy consumption patterns for maximizing their revenue, or by malicious adversaries to derive the living patterns and conduct further intended attacks. For the sake of customers' privacy, personal data and consumption data in smart metering should be protected from unauthorized sharing, disclosing or selling.

On the other hand, one major function of the smart grid system is to collect precise energy consumption data from residential loads and smart meters so that a detailed view of energy usage will be provided to both utilities and consumers. A multitude of energy services are anticipated to be incorporated into the smart grid system to provide value-added services such as dynamic billing, load monitoring and forecasting, demand response, outage and fraud detection, etc. To facilitate such applications, high resolution energy consumption data is expected to be shared among various organizations in the industry and the governments, i.e., between the utilities and third-party service providers, which require access to the metering data at different levels of spatial and temporal aggregation.

While several privacy enhancing techniques (PET) are proposed recently, many of them are based on assumptions that either require extra hardware [2] or are computationally demanding [3], [4]. In this paper, we present a set of solutions for privacy-preserving smart meter data sharing. We argue that practical solutions should be well in line with smart grid data management requirements raised by various relevant stakeholders, particularly taking smart operations and data analytics requirements into considerations. First, we observed that un-anonymized and un-sanitized high-frequency usage data is collected at utilities, which is the *status quo* of smart grid data management in the industry. Although users may proactively manipulate their usage information (e.g., by installing a battery in the household), the high hardware cost may discourage widespread adoption of such mechanisms. Meanwhile, collecting accurate usage data is essential for smart operations (smart consumption, distribution and generation), hence, it is the utilities' best interest to collect original high-frequency usage data and share with its partners in the smart grid system,

978-1-4799-4934-2/14/$31.00 ©2014 IEEE

such as the generators, distributors and marketers.

Therefore, our goal is to prevent such stakeholders from obtaining identifiable smart metering data, while still enabling them to perform their respective functions. In particular, we propose two categories of solutions: (1) we first analyze the privacy threats in the currently published smart metering data sets, and introduce a data sanitization-based mechanism to protect sensitive information before sharing it for external usage; (2) we then present solutions based on secure multi-party computing to enable the third parties to perform aggregation operations on the smart metering data in a privacy-preserving manner. To demonstrate the effectiveness of the proposed solutions, we have implemented our mechanisms and performed experiments on real-world power consumption data.

II. Background and Motivation

System model. Smart meters are installed at each household and connected to the supplier through AMI. The supplier is either an actual power generator or a grid operator (e.g., a utility) with legitimate interest and privilege to collect the identifiable consumption data. We assume there exist multiple third-party data consumers who are allowed to access smart metering data non-intrusively. Hence, the system model is characterized by smart meters, a supplier, and a set of third-party energy service providers. The supplier collects terabytes of fine granular energy consumption measurements stemming from various consumer households, and provides access to the energy service providers in two ways: (1) publishing pseudonymized consumption data to external stakeholders for secondary use; or (2) answering queries of providers.

Data model. Energy consumption data is essentially a collection of temporal sequences, also known as *energy consumption traces*. It consists of records of electric consumption data collected at each household at discrete time slots of equal length. The resolution of a consumption trace is the number of time slots per day. For example, the ISSDA CER Smart Metering dataset [5] studied in this work contains data collected from over 5,000 Irish homes and business participants with a resolution of 48 (i.e., half hourly). The dataset $\mathbf{D}$ consists of m time series records $\{r_m\}$, described by d attributes $A = \{A_1, A_2, ..., A_d\}$, where each attribute A_i is the consumption data observed at time slot T_i. To protect end users' privacy, identifiable information about the inhabitants or households, such as account number, address, phone number, etc., is removed from consumption traces. In industrial practice, they are replaced by pseudonyms [6]. The resulting dataset is considered "anonymized" (or more precisely "de-identified"). An example of the pseudonymized consumption trace is shown in Figure 1, which contains 1-day consumption data of 10 households' at a resolution of 48.

Trust model. Metering data is collected by smart meters at each household and transmitted to suppliers' data server through AMI. We assume all participants (e.g., smart meters, neighborhood collectors/gateways) in this process follow the protocol properly. The supplier is fully trusted to collect and store high-frequency consumption data, while the third-party energy service providers are not trusted to access identifiable energy consumption traces directly. Instead, they are only allowed to access pseudonymized consumption traces or the aggregate values by querying the supplier's data server. In the latter case, we adopt the *honest-but-curious adversary model* (a.k.a. semi-honest model) to describe their behavior. In this model, all parties are assumed to follow the protocol properly ("honest"). Meanwhile, they may keep other parties' inputs and/or intermediate computing results, and actively manipulate these information to infer personal information about others ("curious"). Honest-but-curious adversaries keep the system functioning properly so that they are not identified by intrusion/abnormal detection mechanisms.

MID	T_1	T_2	T_3	T_4	T_5	T_6	...	T_{48}
1	1.212	1.028	0.965	0.876	0.96	0.758	...	1.254
2	0.073	0.067	0.047	0.075	0.069	0.088	...	0.064
3	0.985	1.006	0.949	0.955	0.996	0.956	...	0.926
4	0.192	0.147	0.201	0.141	0.206	0.14	...	0.162
5	0.378	0.363	0.355	0.425	0.304	0.189	...	0.349
6	0.257	0.157	0.298	0.299	0.252	0.25	...	0.862
7	0.481	0.34	0.378	0.276	0.204	0.202	...	0.131
8	0.063	0.129	0.103	0.061	0.127	0.095	...	0.125
9	0.2	0.215	0.264	0.246	0.228	0.225	...	0.215
10	0.395	0.417	0.362	0.338	0.385	0.598	...	0.602

MID	T_1	T_2	T_3	T_4	T_5	T_6	...	T_{48}
1	4	4	3	3	3	3	...	4
2	1	1	1	1	1	1	...	1
3	3	4	3	3	3	3	...	3
4	1	1	2	1	2	1	...	1
5	2	2	2	3	2	1	...	2
6	2	1	2	2	2	2	...	3
7	3	2	2	2	2	2	...	1
8	1	1	1	1	1	1	...	1
9	2	2	2	2	2	2	...	2
10	2	3	2	2	2	3	...	3

Fig. 1. One day consumption traces of 10 households:the left table shows the exact consumptions in kW; in the right table, actual readings are replaced with attacker-defined consumption levels.

The remaining privacy threats. There are primarily two means for the supplier to provide access to consumption traces to other energy service providers. First, after removing identifiable information, the supplier provides the pseudonymized consumption traces to external stakeholders, such as contractors of the supplier providing analysis services or energy service providers making secondary usage of the metering data. However, recent studies show that time series attributes may become *quasi-identifiers* from which an adversary can infer a customer' identity [7] and undo pseudonymization. The goal of the adversary is to link the identity of a target user to his energy consumption trace by customer-specific behavior patterns and unusual energy events (i.e., "behavior anomaly"). A successful attack will allow all the existing deduction attacks to be applied in smart metering. [7] requires the attacker to observe unusual physical events occurring at the target household and attribute consumption traces to individuals based on the rarity of an anomaly. In this work, we further relax the requirements of auxiliary information on the rare discriminative event for the attackers. We assume the attacker can observe the activities of a target household, and classify its energy consumption into l levels based on a variety of factors such as the number of people at home, the number of lights switched on, the charging of an electric vehicle, temperature, etc. The proposed attack and analysis show that even with the relaxed auxiliary information, the pseudonymization consumption traces are at risk of being attributed to individuals. As a result, enhanced PET technologies such as k-anonymity based anonymization and sanitization are needed to protect users' privacy in this type of data sharing in smart metering. In this way, users' privacy expectations are satisfied when consumption data is released or shared externally.

Second, in certain applications, energy service providers need to access raw consumption records to perform (temporal or spatial) aggregate operations. This introduces privacy concerns at both the supplier and the third-party providers. For the supplier, the metering data is considered proprietary not only from the privacy protection requirement but also from business operation perspective. It is impractical to release the collected raw data directly to other parties. Meanwhile, third-party provider rely on the supplier to provide accurate consumption data for secondary usage. To meet the needs, the supplier provides on-demand access by answering aggregate queries from legitimate external parties. On the other hand, third-party providers also want to hide their intentions/business interests from being learned by the supplier. In this paper, we develop a set of privacy-preserving operations to allow third-party providers to privately retrieve consumption data from the supplier with as small as possible communication complexity.

III. Preliminaries and Related Works

Privacy issues with smart metering data. The primary concern appears to be the *inference attack* – AMI data could be utilized to infer behaviors, habits and events in the household, e.g., [8], [9] etc. This category of attacks, also known as *profiling attacks*, mostly employ rule-based, ontology-based [8], supervised/unsupervised learning [9], [10], or information theoretic [11] approaches. User profiling could be launched at various layers: (i) utilizing the electrical circuit features at physical layer – the conventional NIALM [10], [12]; (ii) observing/eavesdropping of device status and control data, metering data at smart grid communication layer [13]; (iii) employing temporal electricity usage data at data management layer [9], or combination of multiple types of data [8]. Countermeasures have been proposed for privacy protection. Most of the existing solutions aim to prevent identifiable, raw high-frequency data from been collected or accessed by untrusted parties.

Privacy issues with AMI data collection process. Undesired information disclosure to the honest-but-curious parties during the AMI data collection process is also considered as privacy breaches. For instance, several approaches have been proposed to employ secure multi-party computation and cryptographical methods to allow intermediate parties to perform operations without accessing raw usage data from other peers. Meanwhile, privacy-preserving metering/billing have been introduced to enable time-of-use billing without collecting raw usage data [14], [15]. The correctness of billing could be proved from technical perspective, however, it remains questionable whether the utilities and the customers would accept this solution.

Homomorphic encryption. Homomorphic encryption represents a group of semantically-secure public/private key encryption methods, in which certain algebraic operations on plaintext can be performed with cipher. Mathematically, given a homomorphic encryption scheme $\mathsf{E}()$, ciphertext $\mathsf{E}(x)$ and $\mathsf{E}(y)$, we are able to compute $\mathsf{E}(x \star y)$ without decryption, i.e. without knowing the plaintext or private keys. $\star$ represents an arithmetic operation such as addition or multiplication. Well-known homomorphic encryption schemes include: RSA, ElGamal [16], Paillier [17], Boneh-Goh-Nissim [18], and etc. The Paillier cryptosystem [17] is additively homomorphic; the El Gamal [16] cryptosystem is multiplicatively homomorphic; and the Boneh-Goh-Nissim cryptosystem approach [18] supports one multiplication between unlimited number of additions. More recent approaches provide full support of both addition and multiplication at higher computation costs [19], [20]. We omit further mathematical details in this paper, since they are out of our scope.

Secure multi-party computing. The original problem of *secure two/multi-party computation* was introduced in [21]. In this problem, multiple parties compute the value of a public function on private variables, without revealing the values of the variables to each other. Zero-knowledge proof [22] addresses the problem of proving the veracity of a statement to other parties without revealing anything else. They are the earliest ancestors of privacy preserving multi-party computing.

IV. Privacy-preserving Data Sharing in Smart Grid Systems

A. Data anonymization and sanitization

In this section, we first introduce a level-based approach that can be used by the adversary to relax the accuracy requirement of auxiliary information needed for re-identification attacks. Then, we present a model based on information gain to theoretically assess the risk of such attacks.

To pseudonymize a smart metering dataset, identifiable information of each contains consumption trace is removed and stored separately in another database. In the pseudonymized consumption dataset $\mathbf{D}$, each record contains three parts: (1) a *pseudonym ID*; (2) a set of *quasi-identifier attributes* at d different time instants, denoted as A_1,..., A_d; and (3) a set of *sensitive attributes*, denoted as A_S as an entirety [23], [24]. In the smart metering setting, the quasi-identifiers are energy consumption data at a selected set of time instants, and the sensitive attributes are energy consumption data of (a subset of) the following time instants that is of interest to the adversary. For instance, in preparing for a burglary, the burglar needs to monitor the victim physically for a long time to obtain a repetitive living pattern of the target to derive the best time for burglary. With a successful re-identification attack, a burglar can easily obtain statistics of the target's long-term energy consumption behavior to derive the living pattern and save a lot of effort in spying on the target.

It is believed that an attacker is capable of performing privacy-invading inference attacks only if he has knowledge of both databases. Hence, releasing pseudonymized consumption dataset is considered secure. However, from our preliminary exploration with the ISSDA CER Smart Metering dataset, we have discovered that unique usage patterns are almost ubiquitous. Therefore, it is possible for an attacker to associate time series consumption data with offline auxiliary knowledge, and effectively attribute consumption traces to individual users/households.

Example 1: *When a burglar observes two households in a same neighborhood, he finds that people at household 1 left for gym at 8:30 pm for half an hour, while household 2 did laundry around 7:00 pm. From home appliance energy usage data [25], we know that 1-hour launch (30-minute washer and 30-minute dryer) introduce 1.9kW on average for household 2, while turning off a plasma TV, two computers and five incandescents for half an hour may save household 1 0.6kW and cause a trough in its evening energy usage. Assume the burglar obtains three consumption traces at a resolution of 48 (i.e., sampling every 30 minutes) as shown in 2, his observations of two households provide ample auxiliary information for him to link consumption trace 1 and 3 to households 1 and 2, respectively.*

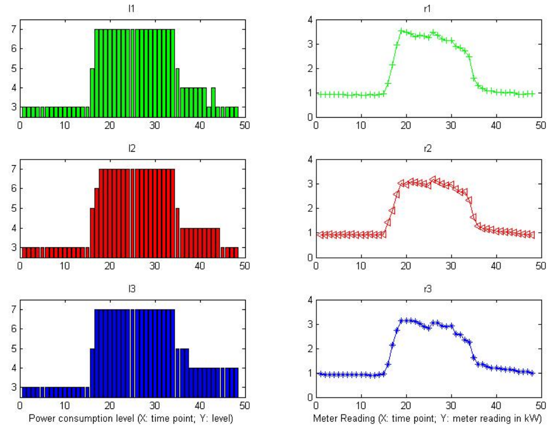


Fig. 2. 100-day average consumption traces of one household reflect level-based consumption patterns.

To understand the potential risks in a published/shared pseudonymized consumption dataset, we first take an entropy-based approach to measure the discriminative information carried in the quasi-identifier attributes. In particular, we denote the known consumption trace of an end user by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where $\mathbf{x}_i$ is the measurements of day i. With a resolution of d, we represent each $\mathbf{x}_i$ as a vector of d dimensions such that $\mathbf{x}_i = \{x_{i_1}, x_{i_2}, ..., x_{i_d}\}$. Each dimension or a combination of dimensions of $\mathbf{x}_i$ could become the quasi-identifier attributes.

Re-identification attacks. Attackers can obtain knowledge about the values of quasi-identifiers from offline channels such as physically observing the activities (leaving or returning to home, the number of people at home) and power consumption indicators (lights switched on/off, smart car charged, temperature, etc.) of the target household. The auxiliary information can be compared with publicly available statistical data to determine an *energy consumption level* E_{l_i} of the target household at time T_i. As an example, a coarsely defined energy consumption level may contain three levels: high, medium, and low. As shown in Figure 2, the average consumption (three traces of 100 days) of one household demonstrates clear consumption patterns with matching levels. To improve the success rate, the attacker should define the consumption levels as fine-grained as possible to provide discriminative indication for identification, however, the increasing granularity tends to be error-prone. We suggest a reasonable value for the consumption level should be between 3 and 10 (for example, in our experiment the consumption is classified into 7 levels). Then, the attacker replaces the data in the consumption traces of each household that fall into the range of level E_{l_i} with its level number l_i. For example, the original consumption dataset (left table in Figure 1) is now transformed into level-based consumptions (right table in Figure 1).

Assessing risk using information gain. To quantify the amount of information provided by an attribute (or an attribute set), we analyze the problem from an information gain perspective. As the goal of the attacker is to attribute a particular consumption trace to an individual, without any prior knowledge, the attacker considers all the traces equally likely to be linked to the target. The average amount of information needed by the attacker (i.e., the expected information gain for a successful identification) is: $E(I(\mathbf{X})) = H(\mathbf{X}) = -log_2(1/n)$ for n households. When the attacker detects the value of attribute A_i as v, the conditional entropy is calculated as: $H(\mathbf{X}|A_i = v) = -log_2(1/N_{A_i=v})$, where $N_{A_i=v}$ is the number of traces whose attribute A_i has a value of v. To assess the actual information gain of knowing an Attribute A_i, we can calculate: $I(\mathbf{X}; A_i) = H(\mathbf{X}) - \sum_{v \in \mathbf{V}_A}(p(A_i = v)H(\mathbf{X}|A_i = v))$. In a dataset of no privacy risk, $I(\mathbf{X}; A_i)$ should be 0. In a poorly anonymized dataset (as the one in Figure 1), the information gain from knowing an attribute is $H(\mathbf{X})$ and this attribute is considered a quasi-identifier. In the CER smart metering dataset, the consumption traces have a resolution of 48 (half hourly). Each of the 48 attributes may be used as quasi-identifier attributes. We studied a dataset of ten households on day $\mathbf{x}_i$ and measured the information gain for each attribute, as shown in Figure 3.

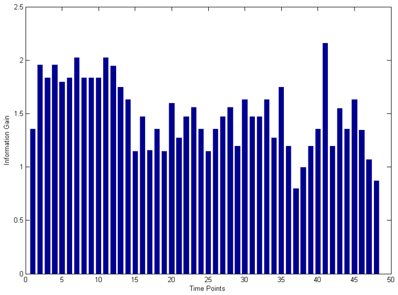
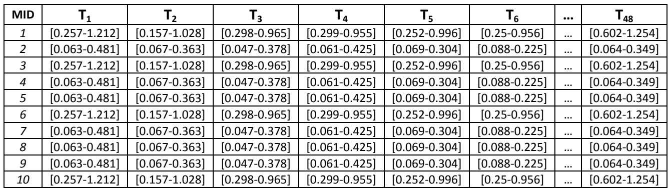


Fig. 3. Information gain with a single attribute.

By defining the consumption level, we release the requirements for attacker's external knowledge, however, it also reduces the chance of successful re-identification. The attacker can employ multiple attributes (consecutive or not) in the attack. For example, when the attacker knows the value of A_1 and A_2, the information gain is denoted as $I(\mathbf{X}; A_1, A_2) = H(\mathbf{X}) - H(\mathbf{X}|A_1, A_2)$. In the current step, we do not consider time patterns in consecutive attributes, therefore, when A_1 and A_2 are independent, we can further represent $I(\mathbf{X}; A_1, A_2) = H(\mathbf{X}) - H(\mathbf{X}|A_1) - H(\mathbf{X}|A_2)$. Intuitively, the more attributes are included, the more likely the attribute set becomes a quasi-identifier attribute set. As the attribute dimension is bounded by resolution, the more fine-grained data, the higher the risk it is exposed to.

Data anonymization. We adopt the concept of k-anonymity [24], [26] to handle temporal data so that records are placed into equivalent groups with at least k consumption traces. Thus, a dataset $\mathbf{D}$ is defined as "k-anonymized" if any target consumption trace T in $\mathbf{D}$ with quasi-identifier attribute set $\{A_1, ..., A_r\} \subseteq \{A_1, ..., A_d\}$, where $\{A_1, ..., A_r\} = \{v_1, ..., v_r\}$, cannot be distinguished from other $k-1$ consumption traces. To anonymize the dataset, quasi-identifier attributes are considered as a vector and mapped to the high-dimensional feature space. Each consumption trace is represented as a node in the r-dimensional space. Here we give an example in Figure 4 to fulfill 5-anonymized for the original dataset in Figure 1.



MID	T_1	T_2	T_3	T_4	T_5	T_6	...	T_{48}
1	[0.257-1.212]	[0.157-1.028]	[0.298-0.965]	[0.299-0.955]	[0.252-0.996]	[0.25-0.956]	...	[0.602-1.254]
2	[0.063-0.481]	[0.067-0.363]	[0.047-0.378]	[0.061-0.425]	[0.069-0.304]	[0.088-0.225]	...	[0.064-0.349]
3	[0.257-1.212]	[0.157-1.028]	[0.298-0.965]	[0.299-0.955]	[0.252-0.996]	[0.25-0.956]	...	[0.602-1.254]
4	[0.063-0.481]	[0.067-0.363]	[0.047-0.378]	[0.061-0.425]	[0.069-0.304]	[0.088-0.225]	...	[0.064-0.349]
5	[0.063-0.481]	[0.067-0.363]	[0.047-0.378]	[0.061-0.425]	[0.069-0.304]	[0.088-0.225]	...	[0.064-0.349]
6	[0.257-1.212]	[0.157-1.028]	[0.298-0.965]	[0.299-0.955]	[0.252-0.996]	[0.25-0.956]	...	[0.602-1.254]
7	[0.063-0.481]	[0.067-0.363]	[0.047-0.378]	[0.061-0.425]	[0.069-0.304]	[0.088-0.225]	...	[0.064-0.349]
8	[0.063-0.481]	[0.067-0.363]	[0.047-0.378]	[0.061-0.425]	[0.069-0.304]	[0.088-0.225]	...	[0.064-0.349]
9	[0.063-0.481]	[0.067-0.363]	[0.047-0.378]	[0.061-0.425]	[0.069-0.304]	[0.088-0.225]	...	[0.064-0.349]
10	[0.257-1.212]	[0.157-1.028]	[0.298-0.965]	[0.299-0.955]	[0.252-0.996]	[0.25-0.956]	...	[0.602-1.254]

Fig. 4. 5-anonymized consumption traces.

Example 2: *To achieve 5-anonimity, we first adopt the k-means clustering algorithm to partition 10 traces into 2 clusters based on Euclidean distance. The number of clusters is determined by the degree of anonymization, e.g., at least 5 traces are needed in each cluster in this example. Then, the numerical values of each attribute in the consumption data is generalized into range values, e.g.,* $0.132kW \rightarrow [0.130, 0.135)$ *kW. Finally, we select the minimal and maximal value of each attribute to generate the range.*

Discussions. Conventional *k*-anonymity approaches generalize data into value ranges to yield k similar records in each group. Different from conventional *k*-anonymity algorithms, we aim to find a balance in generalization of two dimensions, data values and indexes. Generalization should also be applied on sequence indexes by reducing data resolution to covert high-frequency data into lower frequency. It should be application-driven based on the needs of the applications. Pattern is critical for time-series data's users. For example, a electrical company can determine the best time to maintain a user's electrical devices by studying his power consumption pattern. However, the conventional k-anonymity suffers significant pattern loss, so we will apply so-called (k,p)-anonymization which was proposed in [24] to consumption traces to guarantee pattern-preserving anonymity, for example, at least p traces have the same pattern in a group with k indistinguishable members.

B. Secure multi-party computing

As discussed previously, in certain scenarios, the supplier does not share any raw data with external stakeholders. For instance, the utility is prohibited by privacy regulations or service agreements from sharing any type of raw data with external parties, or a third-party application requires to execute aggregate operations on the exact data than on the sanitized data. To address such needs, the supplier can provide on-demand access to third parties, who are allowed to submit temporal or spatial aggregate queries without violating privacy requirements. However, third-party service providers may not want to describe to the supplier the details of its usage of the data (e.g., scope of interest, proprietary algorithms/models for data processing). This requires the supplier to support private information retrieval in the database. In particular, we consider two types of applications: (1) the data processing model/algorithm at the external collaborator is public, while parameters in the model are private; (2) both the data processing model and the parameters are private.

We extend secure multi-party computation techniques [27]–[29] to realize *private function evaluations* at the supplier's database (denoted as "server") for third-party applications (denoted as "client"). In particular, each client defines its own private function F, in the form of $F(\mathbf{s}, \mathbf{Y}) = (s_1 \cdot H(Y_1)) \bigoplus ... \bigoplus (s_i \cdot H(Y_i))$, applying homomorphic operations on the results of $H(\cdot)$. $\mathbf{s}$ is a parameter vector with binary entries encrypted with the homomorphic public key of the client. The client includes both valid operations and dummy operations in function F so that the server cannot infer the right form of F. The parameter for the dummy operations is an encrypted "0", so the results of the dummy operations will not be included in the final result of F. In this way, the privacy of the client is protected.

To protect the privacy of the energy consumer, the server needs to carefully define the input elements Y to ensure that operations $H(\cdot)$ on Y will not violate the privacy requirements. In this work, we consider to support temporal and spatial aggregate queries, and define atomic input element as aggregate consumption of multiple meters or aggregate consumption of a same meter within a time interval. In particular, the server prepares metadata, as shown in Figure 5, to define valid element input Y. In the meta-data, raw consumption data is modeled into a matrix U: each element U_{H_i,t_j} represents the consumption data of household H_i at time t_j. A 2D-1D geographic mapping function $f : (\phi, \lambda) \rightarrow U$ is defined to convert 2D geographic ranges into a set of household indexes $\mathbf{H} = \{H_i\}$, which are located in the given geographic range. Therefore, $\mathbf{U} = \{U_{H_i,t_j}\} : H_i \in \mathbf{H}, t_j \in \mathbf{T}$, where $\mathbf{T}$ is time intervals. Here, we simply define $Y = \sum_{U_{H_i,t_j} \in \mathbf{U}} (U_{H_i,t_j})$. It can be extended to support more complex operations within $\mathbf{U}$ with additional privacy verification. Each element denotes a set of smart meters. The server can define element as large as smart meters in a zipcode, or as small as the smart meter of a building. The server is responsible for defining elements so that the combinations of these elements can satisfy all potential queries from the clients.

Next, we discuss the common operations ($H(\cdot)$) currently supported by our scheme. We define and support seven operations including *privacy-preserving selection*, *weighted selection*, *simple sum*, *weighted sum*, *mean* and *variance*, and will extend to support more in the future.

Privacy-preserving selection. In this operation, H is a selection statement to select households in $\{Y_{i_n,t_j}\}$ and request consumption data of each household at time t_j. To hide the parameters of the private function F, third-party application generates its own S. Once the server receives the query with encrypted parameters, it cannot tell if an element $\{Y_{i_n,t_j}\}$ is actually requested or just a dummy element. Then, the server follows the statement in H to obtain numerical values and conduct the operation in F to get a reply in ciphertext form. Finally, the application recover the result with its private key. As the operation at the server is in ciphertext form, the selection is unknown to the supplier. In summary, this operation takes two communication steps, n homomorphic multiplications at the server, and n homographic encryptions and 1 homographic decryption at the application server. We did simulations to evaluate the overhead of the proposed scheme, and on average it will take the application server 2.48 ms for each encryption and 2.75 ms for each decryption operation, and it will take the data server 8.55 ms for the aggregation of 100 results.

Weighted selection. Similar as privacy-preserving selection, per-range weights can be supported with an application-

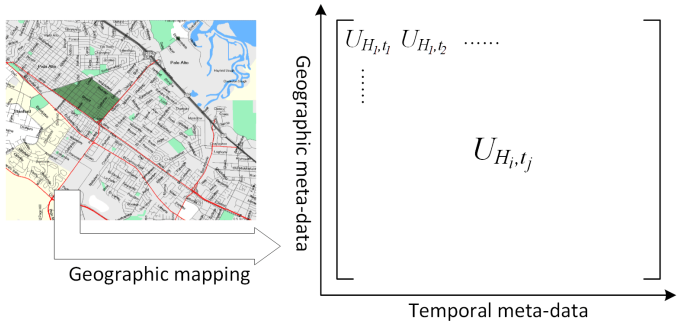


Fig. 5. Data preparation and meta-data publishing.

generated weight matrix $W_{i,j}$. Each entry $w_{i,j}$ should be an integer.

Simple sum, mean & variance. Simple sum is defined as $\sum s_i \cdot Y_{i,t_j}$. From simple sum, *mean* can be derived in two steps, such that external applications get the sum first, and divide by number of households in the selection. Similarly, *variance* is supported in two steps: first sum so that external application gets the mean; then perform $\sum (S_i \times Y_{i,t_j} - m) \times (S_i \times Y_{i,t_j} - m)$.

More operations, e.g., projection, could be supported if they are allowed by the homomorphic cryptosystem, but we do not discuss in the paper due to length limit.

V. Conclusion

In smart grid data management, high-frequency usage data collected by AMI often contains sensitive information about the end consumers. When such data is shared by the utilities with external stakeholders, consumer privacy is at risk. In this paper, we present a set of comprehensive solutions for privacy-preserving smart data sharing. In particular, we prevent external stakeholders from obtaining identifiable consumption data, while still enabling them to perform their respective functions. We have presented solutions based on data sanitization, as well as solutions based on secure multi-party computing. We have performed experiments with real-world energy consumption data. The results show that the proposed solutions are both effective and efficient.

VI. Acknowledgment

This work was partially supported by NSF under award NSF0073319, EPS0903806 and matching support from the State of Kansas through the Kansas Board of Regents.

References

[1] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter," in *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*. ACM, 2010, pp. 61–66.

[2] G. Kalogridis, R. Cepeda, S. Denic, T. Lewis, and C. Efthymiou, "Elecprivacy: Evaluating the privacy protection of electricity management algorithms," *IEEE Trans. on Smart Grid*, vol. 2, no. 4, pp. 750–758.

[3] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Proc. STM'11*, 2011, pp. 226–238.

[4] C. Rottondi, G. Verticale, and A. Capone, "A security framework for smart metering with multiple data consumers," in *INFOCOM Workshops*, march 2012, pp. 103 –108.

[5] ISSDA, "CER smart metering project: Electricity customer behaviour trial," http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[6] J.-M. Bohli, O. Ugus, and C. Sorge, "A privacy model for smart metering," in *Proceedings of the First IEEE Workshop on Smart Grid Communications (in conjunction with ICC 2010)*, 2010.

[7] M. Jawurek, M. Johns, and K. Rieck, "Smart metering de-pseudonymization," in *ACSAC*, 2011, pp. 227–236.

[8] H. S. Cho, T. Yamazaki, and M. Hahn, "Aero: extraction of user's activities from electric power consumption data," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 2011–2018, 2010.

[9] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *Security & Privacy, IEEE*, vol. 8, no. 1, pp. 11–20, 2010.

[10] M. Zeifman, "Disaggregation of home energy display data using probabilistic approach," *IEEE Trans. on Consumer Electronics*, vol. 58, no. 1, pp. 23–31, 2012.

[11] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor, "Smart meter privacy: A utility-privacy framework," in *IEEE SmartGridComm*, 2011, pp. 190–195.

[12] M. Zeifman and K. Roth, "Nonintrusive appliance load monitoring: Review and outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 76–84, 2011.

[13] D. Li, Z. Aung, J. Williams, and A. Sanchez, "P3: Privacy preservation protocol for appliance control application," in *IEEE SmartGridComm*. IEEE, 2012, pp. 294–299.

[14] A. Rial and G. Danezis, "Privacy-preserving smart metering," in *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*. ACM, 2011, pp. 49–60.

[15] C.-I. Fan, S.-Y. Huang, and W. Artan, "Design and implementation of privacy preserving billing protocol for smart grid," *The Journal of Supercomputing*, pp. 1–22, 2013.

[16] T. E. Gamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," in *CRYPTO*, 1985, pp. 10–18.

[17] P. Paillier, "Public-key cryptosystem based on composite degree residuosity classes," in *Proceedings of Eurocrypt '99*. Springer-Verlag, 1999, pp. 223–238.

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

[19] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *STOC*. New York, NY, USA: ACM, 2009, pp. 169–178.

[20] M. van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully homomorphic encryption over the integers," in *EUROCRYPT'10*, 2010.

[21] A. C. Yao, "Protocols for secure computations," in *IEEE SFCS*. Washington, DC, USA: IEEE Computer Society, 1982, pp. 160–164.

[22] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof-systems," in *STOC '85*, 1985, pp. 291–304.

[23] R. G. Pensa, A. Monreale, F. Pinelli, and D. Pedreschi, "Pattern-preserving k-anonymization of sequences and its application to mobility data mining," *Privacy in Location-Based Applications*, 2008.

[24] X. Shang, K. Chen, L. Shou, G. Chen, and T. Hu, "(k, p)-anonymity: towards pattern-preserving anonymity of time-series data," in *CIKM*, 2010, pp. 1333–1336.

[25] G. Electric, "How much power do your appliances use?" http://visualization.geblogs.com/visualization/appliance.

[26] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, 2002.

[27] V. Kolesnikov and T. Schneider, "A practical universal circuit construction and secure evaluation of private functions," in *Financial Cryptography and Data Security*. Springer, 2008, pp. 83–97.

[28] R. Canetti, Y. Ishai, R. Kumar, M. K. Reiter, R. Rubinfeld, and R. N. Wright, "Selective private function evaluation with applications to private statistics," in *PODC*, vol. 1, 2001, pp. 293–304.

[29] A. Paus, A.-R. Sadeghi, and T. Schneider, "Practical secure evaluation of semi-private functions," in *Applied Cryptography and Network Security*. Springer, 2009, pp. 89–106.