Privacy-Aware Profiling and Statistical Data Extraction for Smart Sustainable Energy Systems

Huang Lin and Yuguang Fang

Abstract—**The growing population and global warming have been calling for more effective energy usage, which have stimulated the emergence of smart sustainable energy technology. The distinct feature of this newly emerging technology is the incorporation of advanced information and communication technologies (ICT), which collects more detailed information on how energy is generated, distributed, and consumed. Various smart metering technologies have also been proposed to support the optimization on sustainable energy usage. Despite the obvious benefits of these technologies, people may still hesitate to adopt them because of possible privacy breach. On the other hand, we observe that the major target information for making the sustainable energy system smart is the aggregated statistics of energy usage, not the full detailed usage profiles which would compromise customers' privacy. Thus, how to design schemes to collect aggregated statistics while preserving customers' privacy becomes an important research problem. In this paper, we propose two schemes to deal with this problem. The first one can support dynamic profiling, which can extract aggregated statistical information without compromising individual privacy. The second one aims to extract correlation information among various factors for the smart system design and can also be used as an underlying tool for baseline inference and association rule mining.**

Index Terms—**Privacy, smart grid, smart sustainable energy, dynamic profiling.**

I. Introduction

SUSTAINABLE energy technology has attracted a great deal of attention due to the tremendous population growth and global warming. Various novel concepts such as "smart grid," "smart home," "smart city," etc., have been proposed. A notable feature of the newly emerging "smart" sustainable energy technology is the incorporation of advanced information and communication technologies (ICT). In the newly proposed smart grid, a novel communication network supervisory control and data acquisition (SCADA) has been proposed to facilitate more fine-grained information collection on energy consumption, and exert more control on the individual energy usage. The wide deployment of low cost portable sensing and metering devices and the advancements of mobile smartphone technologies will further accelerate the incorporation of ICT in developing efficient sustainable energy systems. It is envisioned that a large amount of information will be collected and processed by using those metering and mobile computing devices deployed in the sustainable energy usage systems, such as home energy systems, microgrids, vehicular storage systems, etc. The collected information could be either related to detailed individual energy consumption patterns [1], [2], or information on individual behavioral patterns such as the detailed position or speed of an individual vehicle [3], [4], or home energy consumption habits and profiles [2]. Obviously, the ICT can not only help regularly and interactively provide profiling and forecasting for effective energy generation and distribution to optimize the demand and response matching in the short term, but also help researchers and future designers of sustainable energy systems to better understand either the pros and cons of the current city planning systems [4] or the psychological and behavioral rationales behind individual energy usage patterns [2], [5], [6], which could further result in more efficient sustainable energy system design methodologies.

However, despite the potential benefits of these technologies, people are still hesitating in adopting them due to the concern of privacy leakage [7]. The worry about losing privacy is partly due to the fact that the customers do not know in what manner the data aggregator deals with the collected individual data profile for energy usage. It has been noted that the target collected information useful for efficient sustainable energy system design is mostly about the aggregated statistical data, such as the aggregated or average data, which can already provide solid sufficient information for making good decision and improving the effectiveness of sustainable energy systems. For instance, the knowledge on the general correlation between the household temperature and the customer energy consumption is enough for a regional manager of a smart grid to capture the energy consumption fluctuation due to the temperature change [8]. To enable the designers of a sustainable energy system to better understand the needs, attitudes, motives and behaviors of energy consumers, researchers [9], [10] have already proposed various systems in collecting statistical information on the relationship between the social characteristics such as socio-economic structure, and the relevant energy usage patterns. For instance, researchers have attempted to identify the correlation between the customers' social background [9], such as income, age, educational background, family structure, household size, race, user skills, individual habits and political beliefs and social values or norms the customers cherish and the level of "sustainable life style" they adopt. A research conducted in 2006 [11] attempted to figure out the correlation between how users form the energy-consciousness and how people alternate exactly which parts of their homes, such as their bathrooms, or bedrooms, etc., when they move into their new homes. When an aggregator attempts to understand what kinds of social norms or behavioral patterns are the driving forces for individuals to do whatever they are doing with energy usage, the privacy issues become

Manuscript received April 04, 2012; accepted July 09, 2012. Date of publication September 04, 2012; date of current version February 27, 2013. This work was supported in part by National Science Foundation under Grant CNS-0916391. Paper no. TSG-00172-2012.

The authors are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32601 USA.

Digital Object Identifier 10.1109/TSG.2012.2210289

much more prominent than ever. People are generally reluctant to let others dig into their daily life and figure out what they are actually thinking about without a stringent privacy guarantee [6]. This, combined with the smart metering technology, which has already been deemed as a dangerous privacy breach of one's daily life [12], makes the privacy-aware profiling (or simply private profiling) and forecasting design especially challenging and intriguing.

Since the aggregated statistics are really the target for the design of sustainable energy systems, this paper focuses on providing more effective and efficient technology for dynamic privacy-aware profiling, i.e., collecting the target statistical information without compromising the privacy of the individuals who submit information. In other words, to overcome the "fear roots on ignorance" effect on profiling, the target statistical information should be well defined and publicly known before the profiling starts. People might be more willing to participate in the profiling process provided that they are aware of and have control over what kind of statistical information are leaked to others. There should be even more incentive for individuals to participate and provide honest responses in the profiling especially when the target statistical information benefits them, for example, the statistical information with regard to energy consumption behavioral patterns could help provide educated advices on how they can save their energy consumption at home.

Due to the inherited unpredictability of its variables, the profiling and forecasting for sustainable energy systems are highly dynamic and thus tricky. Indeed, when the grid operators [7] in The Netherlands were asked to define their goals on what kinds of information they request to accomplish their profiling, they found it very difficult to provide a specific definition. The underlying reason is that the optimization of a smart grid sometimes might depend on certain information which is almost impossible to predict such as the room temperature, or the socio-economical background of future residents in a concerned area [1]. Census is one of the most commonly used forms when it comes down to the social or economic research on the customer profiling due to its flexibility and ability to adjust to a highly dynamic system. Using census allows a data aggregator to define the concrete profiling goal on the fly. The questions of a census can be designed independently of the categories of users, which is essentially highly dynamic and impossible to predict.

The existing profiling schemes in sustainable energy systems [1], [7] require the grouping of metering devices according to predetermined categories, which are difficult to adjust to the frequently changing and unpredictable attributes such as customers' ages, locations, time, demographics, or individual behavioral patterns, or political beliefs or dynamic environmental factors such as room temperature or humidity. The current schemes [1], [7] also face challenges when the dynamic grouping of users rely on private customers' social or economical characteristics such as their incomes or locations which could raise serious privacy concerns because they are difficult to gather in the first place, let alone grouping the users according to those data. Besides, even without considering those issues, the rekeying operations for secure information management when changing meter groups are highly costly in the current schemes, and hence could bring significant inconvenience to the customers, which will further lower the users' willingness to participate.

In this paper, we propose two schemes to deal with this problem. Our schemes enable a data aggregator to extract the statistical information without relying on any regrouping operations, and hence are suitable for dynamic profiling. The proposed schemes employ the privacy-aware census (or simply private census) [13] as the underlying tool. The first private census scheme can enable an aggregator to extract the summation information from the submitted individual responses, which could privately answer the statistical question in the form as "What is the total energy consumption when the home temperature is 25 °C?" Since the correlation among various variables is vital for the design of sustainable energy technology [8]–[10], we further propose the second pattern oriented privacy-aware profiling scheme. It intends to efficiently extract the intersection information, which can answer the query in the form as a conjunction "How many more percent of users consume how much energy on average when they satisfy both the following conditions: have an annual income larger than 100 000 dollars **AND** the room temperature is 25 °C?" The latter scheme could further serve as an underlying tool for private baseline inference and data mining association rule when it is combined with the former scheme. Associate rule mining was proposed by Agrawal *et al.* [14], which targets to discover interesting relationships between variables in large databases to facilitate marketers to develop customized marketing strategies [15]. We believe the associate rule mining will benefit the energy policy makers and designers in the long term.

II. Related Work and Our Contribution

Private smart metering. One of the major focuses of research on the privacy issue in smart grid is to design private billing schemes based on the detailed private information collected from smart metering devices. Most of these schemes focus on collecting aggregated energy consumption data from an individual customer without the leakage of the detailed energy consumption information of this consumer. The current private billing schemes intend either to reduce the implementation cost [16] or to introduce extra privacy mechanism such as differential privacy [17], [18] to the billing system. None of the existing schemes consider the privacy-aware profiling issue across different metering devices.

Behavior and smart grid. A thorough survey on the relationship between customer behaviors and the sustainable energy technology can be found in [6]. Researchers have made significant efforts on investigating how customer energy consumption behaviors can be affected by their psychological and socio-economical backgrounds, and how the customer behaviors can in turn affect sustainable energy systems. Albeit privacy issue [6] has been considered imperative when the social profiling meets the sustainable energy technology, very few results can be found in this direction.

Private social profiling for smart grid. The wide range of implementation and deployment of smart grid technology, such as smart metering, expects much more involved users and a better understanding of end users behaviors, which could further serves as a vital feedback for the analysis of smart grid systems. An interesting privacy friendly aggregated protocol for smart grid was proposed by Kursawe *et al.* [1]. The proposed protocol provides a private aggregation protocol for data collected from the smart metering devices. The protocol relies on a predefined grouping of users, which might not be practical for customer profiling due to possible privacy concerns and the essentially dynamical and unpredictable nature of profiling. A statistical linear regression scheme [1] was mentioned to estimate the aggregation data for an arbitrary group. However, the statistical estimation method relies on *a priori* knowledge on some unknown parameters such as the exact number of metering devices belonging to an arbitrary group, which itself might be the target information and thus difficult to obtain in the first place. Besides, the statistical estimation approach based on the aggregation data of all system users might fail to provide estimation for a specific group with enough accuracy due to the limited input information.

Our Contribution. Our protocol avoids relying on the re-grouping of users by simply letting an aggregator publish a private census which can adjust to any kind of investigation purpose, and guarantee the aggregator can only extract the targeted statistical information without any extra individual information and thus reduce the cumbersome communication overhead due to rekeying operations for grouping. To improve the performance of a general census scheme, we provide an improved pattern oriented privacy-aware survey scheme aiming to efficiently extract the choices made by most users simultaneously. The simultaneously chosen answer can serve as an important tool for the system designer to discover the underlying principle that runs the whole sustainable energy system. Aside from being useful in the private profiling in smart grid, the proposed scheme can also be applied to the private information aggregation for other applications [3].

III. System Model

We first present the system model in which our private profiling schemes are developed. Our protocol aims to provide private profiling solution for smart metering assisted sustainable energy systems such as home electricity, water, gas, smart vehicles, etc. We note that smart metering devices could either refer to home smart metering devices or metering devices attached to their appliances, or the on-board unit devices [19] connected to the GPS in vehicles. There are three parties in our system: metering devices, individual users, and aggregators. The user is referring to a user with computing agents such as the user's smart phone, or a software provided by a third party trusted by the user just as in most of the recent private billing systems [20], [16] for smart metering. The protocol guarantees the individual's privacy, which means the aggregator is only allowed to collect statistical information predefined and agreed upon by the involved users before the profiling system is started. The statistical information should not only indicate the specific target statistical information, i.e., such as "How many more percent of users consume how much energy on average when they satisfy both the following conditions: have an annual income larger than 100 000 dollars **AND** the room temperature is 25 °C?," but also how the received information will be used, e.g., whether it will be publicly released or as a reference for policy decision making. The latter information can serve as a reference for a user to decide whether to participate in the profiling or not. The system also provides a mechanism to verify the correctness of users' responses which can be deduced from the metering data. For example, the integrity of a user's response can be guaranteed when he is asked about his energy consumption during a certain time period because the data can be deduced from the detailed metering data.

The system time is divided into fixed time periods. At the beginning of each period, each user, either a new customer or an existing one, registers to the system and obtains a private key by running the secret key distribution protocol to be introduced. The total number of system users is assumed to stay unchanged during a time period. During each time period, the aggregator can publish multiple censuses either simultaneously or separately to each individual user, who will encrypt his response under his private key and return the ciphertext to allow the aggregator to extract the target statistical information. We note that a registered user can either choose to participate or not. Each question in the published census will be identified by a unique index I, which is publicly known to each individual in the system. The aggregator can either release the statistical information or make a decision based on the received information according to the predefined design objective. Compared with the original aggregation protocol [1], the proposed model has the benefit that users do not have to exchange information with each other each time when the aggregator wishes to collect aggregated data. Users only need to communicate with others at the beginning of a time period, and then are only required to deliver the responses to the aggregator for the rest of time period. This will save the users tremendous efforts especially when the census is frequently conducted.

IV. Preliminary: Secret Key Distribution

Before we present our schemes, we first briefly discuss the preliminary materials. At the beginning of each time period, each of the n users gets a private key used for the subsequent private census schemes. We provide two kinds of secret key distribution models: semi-decentralized model and fully decentralized model. These two models correspond to the interactive protocol and the Diffie-Hellman key-exchange based protocol in the original aggregation scheme [1], respectively.

The first secret key distribution scheme is based on secret-sharing scheme and selects ℓ leader users, who are trusted not to collude with each other. The number of leader users $1 \leq \ell \leq n$ is a vital parameter. We note that when $\ell = n$, the proposed secret key distribution scheme can be considered as a joint zero secret sharing scheme [21], which corresponds to the fully decentralized model. The semi-decentralized model corresponds to the case when $\ell < n$. At the beginning of each time period t, each user delivers an encrypted share to the leaders, who then computes the final shares such that all shares sum to zero together. We assume that the public keys $PK_{(k)}$ for the leader users u_k are certified and published to the rest of users just as in [1] such that the share could be safely delivered to the leader users through encryption. The individual user u_i will keep the added shares as his private key SK_i in the specific time period. The private key will be used to mask the users' response and ensure the aggregator can only get the predefined target statistical information. Formally, the secret key distribution scheme runs as follows:

1) To generate masking values, each user u_j computes ℓ random values $s_{j,1}, \ldots, s_{j,\ell}$. It then encrypts $s_{j,k}$ under the public keys for the leaders whose identities are $u_1, \ldots, u_\ell$, respectively (without loss of generality we

assume the first ℓ users are the leaders). The set of ℓ encrypted shares will be sent to the respective leader users. The encrypted shares can also be first transferred to the aggregator and then re-delivered to the leader users as in [1].

2) Each leader u_k collects $n-1$ shares $s_{j,k}, 1 \leq j \leq n, j \neq k$, and computes its own share $s_{k,k}$ such that the summation of all shares is equal to 0, i.e., $\sum_{j=1}^{n} s_{j,k} = 0 \bmod q$, where q is a Diffie-Hellman prime number, which will be further used in the subsequent private census schemes.

3) Finally, each user u_j adds all his shares $s_{j,1}, \ldots, s_{j,\ell}$ as his private key $SK_j = s_{j,1} + \cdots + s_{j,k}$. We note that $\sum_{j=1}^{n} SK_j = 0$ always holds.

V. General Private Census Scheme for Dynamic Profiling

In the general private census scheme, an aggregator publishes a census consisting of N questions $Q_I, I \in [1, N]$, the response to which could be either 0 or 1, or a certain metering data D_I from a certain metering device. The design of the census questions depends on the concrete application scenario. For instance, when the question intended to answer is "How much energy on average is consumed when the house temperature is 25 °C in the next thirty minutes?," this question can be divided into two questions "What is the total energy consumption when the home temperature is 25 °C?" and "How many homes have a temperature at 25 °C?" User u_j can simply deliver the concrete metering data D_j as his answer when his home temperature is indeed set at 25 °C, and 0 otherwise. For the second question, he just answers 1 if his home temperature is set at 25 °C and 0 otherwise.

u_j encrypts his answer (i.e., response) A_I for each question Q_I as follows: for the binary answer $A_j(I) = 1$, or 0, u_j simply encrypts the respective answer as $g^{A_j(I)} \cdot H(I)^{SK_j}$, where H is a public hash function modeled as a random oracle. When the respective answer is a metering data $A_j(I) = D_I$, the encrypted answer forms as $g^{D_I} \cdot H(I)^{SK_j}$. The reason for the "special treatment" of the metering data related answer will be obvious when the user is required to prove the correctness of his answer to be introduced later.

For the first case, the aggregator, upon receiving those encrypted messages, will compute the aggregated encrypted answer as $g^{\sum_j A_j(I)} \cdot H(I)^{\sum_j SK_j} = g^{\sum_j A_j(I)}$. The masking value due to the private key is canceled out since $\sum_j SK_j = 0$ holds. For the binary answer, the aggregator can simply compare $g^{\sum_j A_j(I)}$ with the number in $[g^1, g^n]$. For the answer related to the metering data, the aggregator can still compute the answer for each question by comparing $g^{\sum_j D_I} \cdot H(I)^{\sum_j SK_j} = g^{\sum_j D_I}$ with the number in $[g^{\text{Min}}, g^{n\text{Max}}]$, where Min and Max denote the minimum and maximum individual metering data, respectively. We note that the difference between Min and Max is assumed to be not too large since our major concern is on protecting the aggregated individual detailed energy consumption data [12] in a short time interval, in which the individual privacy is most likely to be breached from his answer if without a protection mechanism. Hence, $\sum_j D_I$ can be efficiently computed by the comparison approach.

The security of the general census scheme can be stated in the following result.

Theorem 1: Assuming that the Decisional Diffie-Hellman assumption holds in group $\mathbb{G}$, and that the hash function H is a random oracle, then the only useful information the aggregator can obtain through the general census scheme is the targeted aggregated statistical information for dynamic profiling.

Proof: Our proposed general census scheme can be viewed as a combination of the underlying secret key distribution scheme and the distributed summation protocol (without the additional noise r) proposed by Shi *et al.* [18]. We guarantee the general census scheme satisfies the "encrypt-once security" condition by using different index I for each question. Therefore, we have $\forall u_j, \forall A_j(I)$ or D_I, the tuple $(u_j, I, A_j(I)(\text{or } D_I))$ would not appear twice in the encryption query. Our general census scheme replaces the trust dealer in the original construction [18] by the secret key distribution scheme. Therefore, the proposed general census scheme can guarantee the only information the aggregator obtains is the targeted summation as long as the underlying secret key distribution is secure and the distributed summation protocol is aggregator oblivious under random oracle. The security of both protocol can be reduced to the decisional Diffie-Hellman assumption. □

This general census scheme can be directly applied to facilitate the dynamic profiling. Take the above "average energy when home temperature is 25 °C in the next 30 minutes" as an example, the existing data aggregation schemes [1] would either require the pre-grouping of users according to the household temperature or using statistical linear regression based on the aggregated data of all n users and the number of homes with a temperature at 25 °C. The former solution is basically infeasible since the re-keying operation would be highly costly. Our solution can provide the exact answer for this type of dynamic profiling issue while the accuracy of the latter estimation approach is poor.

Although the general private census scheme can serve as a powerful tool for dynamic profiling, it still leaves a large amount of information out of the picture. Let us consider a census with three binary questions A, B, and C. A direct application of the general census scheme can tell the aggregator about the number of users choosing 1 for each question. However, compared with the Venn diagram on the right side, we can observe that the information related to the intersection of those questions are left outside of picture by this direct application. One might argue that the aggregator can simply publish four extra questions of conjunction form such as "how many users choose 1 for question A **AND** B simultaneously" or "how many users choose 1 for question A **AND** B **AND** C simultaneously" to cover all the regions in the Venn diagram. However, this is apparently not a practical solution since the question list size could expand exponentially.

We notice that the missing information in the general census scheme is substantial for the study of relationship between various variables in the system. Take the relationship between energy consumption and home temperature as an example. Suppose that the aggregator wishes to find out the most likely energy interval when the home temperature is in a certain interval. A trivial solution using the general private census scheme is shown as follows: the aggregator first divides the energy consumption range $[\text{Min}_1, \text{Max}_1]$ into G_1 intervals $\{[L_{1i}, U_{1i}]\}_{i=1}^{G_1}$, and the temperature range $[\text{Min}_2, \text{Max}_2]$ into G_2 levels $\{[L_{2i}, U_{2i}]\}_{i=1}^{G_2}$.

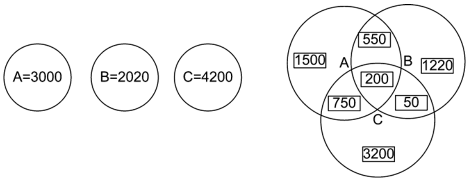

Fig. 1. The loss of statistical information in the general census scheme.

Then the posed question can form as "Is your energy consumption data in $[L_{1i}, U_{1i}], i \in [1, G_1]$ **AND** when the home temperature is $[L_{2j}, U_{2j}], j \in [1, G_2]$?" Apparently, there will be $G_1 \times G_2$ questions in the posed questions. The question list will expand dramatically when extra private or dynamic variables such as income or location, are introduced in the above target question. Even without considering the conjunction question between various variables as introduced in the Venn diagram example, there are already $\prod_i G_i$ where G_i is the fine-grain level for each variable. The size of the question list will expand dramatically when the conjunction questions between different variables are also considered. Since the individual overhead is dependent on the size of the question list, which would render the trivial solution infeasible in the above scenario.

One might argue that the aggregator could simply choose the conjunction questions that he thinks is interested in to exert control on the expansion of the census question list. This leads to another question: how does the aggregator know which questions he should be interested in when he has little or even no *a priori* information on the underlying pattern of the variables involved. In the following section, we provide a solution for the pattern oriented private census scheme aiming to extract the intersection pattern of the majority of users. The information will serve as a reference to help the aggregator to accomplish a more thorough and effective pattern profiling compared with the trivial solution.

VI. Pattern Oriented Census Scheme

The pattern oriented private census scheme aims to provide the aggregator the ability to find out which choices in the question space are simultaneously selected by multiple users in the system. We first present a scheme in which the choices will pop out in the aggregated result when they are simultaneously chosen by all n users. Then we show how this scheme can be transformed into a system in which the intersection of choices can be detected even only more than a threshold number of users select them simultaneously. Our proposed pattern oriented private census scheme is modified from the private information extraction scheme proposed by Lin *et al.* [13]. Compared with the trivial solution which could have excessive overheads, the proposed pattern oriented census scheme only has overhead of linear growth, which is dependent on the summation of the fine-grain levels for all variables. In other words, the individual overhead depends on $\sum_j G_j$.

The basic scheme incorporates the polynomial representation technique from [22] and the general private census scheme together. We take the polynomial representation for the intersection of two sets as an example to introduce the basic idea for polynomial representation technique. Given a set $S_1 = \{a_j\}, S_1$ can be represented as a polynomial $f_1(x) = \prod_{1 \le j \le k} (x - a_j)$ in a polynomial ring $R(x)$ consisting of all polynomials with coefficients from the ring R. In here, a appears in the set S_1 iff $(x - a) | f_1(x)$, where "|" means "divisible." The intersection of two sets $S_1 \cap S_2$ is defined as the set in which each element a that both appear in S_1 and S_2. Let S_1 and S_2 be two sets of equal size, and f_1 and f_2 be their polynomial representations, respectively. The polynomial representation of $S_1 \cap S_2$ as: $f_1 g_1 + f_2 g_2$ where $g_1, g_2 \leftarrow R^{\ge \deg(f_1)}[x]$, where $R^{\ge \deg(f_1)}[x]$ is the set of random polynomials with degree no lower than the degree of f_1 and $\deg(f_1)$ denotes the degree of the polynomial $f_1(x)$. It has been shown in Theorem 2 in [22] that if each player P_i inputs a polynomial f_i representing P_i's set S_i, then the mere information the third party can extract from the polynomial $\sum_{i=1}^{n} f_i g_i$ is the intersection information $S_1 \cap \ldots \cap S_n$, where g_i is a random polynomial with degree $\ge \max_i \deg(f_i)$. In our proposed basic pattern oriented private census scheme, each user represents his choices as a polynomial and then randomizes the polynomial representation with a random polynomial just as in the above polynomial representation technique. Then he can encrypt the resulting polynomial coefficients as in the general private census scheme. The security of the general private census scheme can guarantee that the aggregator can only obtain the summation of the randomized polynomial, which only represents the intersection of choices according to Theorem 2 in [22]. The concrete scheme is formally stated as follows.

There are three phases in our scheme. In the first phase, the aggregator broadcasts the pattern oriented census to users. We provide a toy example here for better illustration and the goal of the aggregator is to find out the correlations among the following few variables: "The average home energy consumption in a certain region," "the income," "age," "education: college or not." The aggregator first divides the range of the above variables into several intervals. For example, he can divide the "age" into 20 intervals as " [15], [19]," " [20], [24]," etc. Here the minimum and maximum are 15 and 114, and the fine-grain level is 20. The census questions consist of all the intervals for those variables and the respective question index I. We note that the variable could also be binary such as the question with regard to the last variable "education: college or not" is binary, and the respective fine-grain level is 2. However, different from the general census scheme, the answer to the binary question should be either I or $-I$, where I is the question index, rather than using 1 or 0 as an answer. The reason is that there could be multiple binary questions and this is the only way to distinguish them in the final aggregated result. The census could also consist of another independent question as "are you willing to participate in this question?," which will be useful in the threshold decision case to be introduced. We notice that all the questions in the census must be attached with a unique index I, and all the indexes should be different from the bounds of the intervals to avoid confusion, which is easy to do.

Let $\mathbb{G}$ denote a cyclic group of prime order p in which decisional Diffie-Hellman (DDH) is hard. Let $H : Z \rightarrow \mathbb{G}$ denote a public hash function. Assuming there are V variables, each of which will be either represented as G_i interval choices

$[L_j, U_j], j \in [1, G_i], i \in [1, V]$ or a binary question. We first deal with the scenario where all the users deliver their real answers.

For question $Q_I, I \in [1, V]$, the answer of user u_j could be either an interval $A_j(I) = [L_j(I), U_j(I)]$ or the respective index $A_j(I) = I$, or $-I$. Without loss of generality, we assume the first V' questions are binary questions and the rest $V - V'$ questions are interval questions. The answer is represented as a set $S_j = \{A_j(I), I \in [1, V'], [L_j(I), U_j(I)], I \in [V' + 1, V]\}$, and then u_j further computes the polynomial $f_i = \prod_{A_j(I) \in S_j, [L_j(I), U_j(I)] \in S_j} (x - A_j(I))(x - L_j(I))(x - U_j(I))$. Then u_j chooses a random $2V - V'$-degree polynomial g_j over the ring and multiplies two polynomials $f_j g_j = \sum_{i=0}^{4V-2V'} c_{ji} x^i$.

Then for each coefficient $c_{ji}, i \in [0, 4V - 2V'], u_j$ computes the following ciphertext $C_{ji} = g^{c_{ji}} H(i)^{SK_j}$, where $i \in [0, 4V - 2V']$. At the end of each round, the individual ciphertext $C_j = \{C_{ji}, i \in [0, 4V - 2V']\}$ is submitted to the aggregator.

At the end of each census, the aggregator collects all the individual ciphertext $C_j = \{C_{ji} = g^{c_{ji}} H(i)^{SK_j}, i \in [0, 4V - 2V']\}$. For each $i \in [0, 4V - 2V']$, it is easy to compute $\prod_{j=1}^{n} g^{c_{ji}} H(i)^{SK_j} = \prod_{j=1}^{n} g^{c_{ji}}$. The target polynomial is $F(x) = \sum_{j=1}^{n} f_j g_j = \sum_{i=0}^{4V-2V'} \sum_{j=1}^{n} c_{ji} x^i = \sum_{i=0}^{4V-2V'} e_i x^i$. $F(x)$ corresponds to the target intersection information. If everyone selects an answer $A(I)$, either it is an index or an interval, we would have $(x - A(I)) | F(x)$. In order to find out which choice or index $A(I)$ in the question list is the intersection of choices, the aggregator only needs to check whether $(x - A(I)) | F(x)$ holds as follows: for each $A(I)$, given $g^{e_j}, j \in [0, 4V - 2V']$, the aggregator checks whether $g^{F(a)} = \prod_{i=0}^{4V-2V'} (g^{e_i})^{A(I)} = 1$ holds. Indeed, $g^{F(a)} = g^0 = 1$ holds if and only if $(x - A(I)) | F(x)$ holds.

The above scheme only reveals to the aggregator the intervals or the indexes which are chosen by all the users. The requirement for the above intersection information might be too stringent. A practical question generally forms as "more than a certain percent of users consume a certain amount of energy on average in a certain region **AND** have an income of certain amount **AND** the age range **AND** have been educated in college." In other words, we need to provide the aggregator about the pattern of the majority users rather than all the users. In the following scheme, we assume there are only k users who will provide their real answers, and the others serve as the shadow users whose sole function is to hide the identities of the users providing real answers among them. The above scheme should be modified as follows.

At the first stage, there is an extra question in the form of "are you willing to participate in this question?" This question is treated as an independent general census question attached to the pattern oriented census scheme. The user who provides the real answer generates the answer set and the respective polynomial representation exactly as in the above scheme. However, u_j should choose a random $5V' + 4\sum_{i=V'+1}^{V} G_i - 2V$-degree polynomial over the ring to master his $2V - V'$ polynomial representation. In other words, the randomized polynomial for each individual is of degree $5V' + 4\sum_{i=V'+1}^{V} G_i - 2V + 2V - V' = 4V' + 4\sum_{i=V'+1}^{V} G_i$. The rest shadow users not providing the real answers simply choose all indexes and the intervals in the question list as his answer (whose polynomial representation is of degree $2V' + 2\sum_{i=V'+1}^{V} G_i$) and choose a random $2V' + 2\sum_{i=V'+1}^{V} G_i$ polynomial as the mask polynomial. Since the answer for those shadow users are basically the whole question space, then the intersection of the answers for all the k users provides the real answer and the answers of the shadow users are basically the intersection answers for the k users. We note for the general census question, those in the k users answer 1, and the others will answer 0. This modified pattern oriented census scheme provides the aggregator a more relaxed version of intersection answer, which only corresponds to the k of n system users.

The security of the proposed pattern oriented private census scheme can be stated as follows.

Theorem 2: Assuming that the decisional Diffie-Hellman (DH) assumption holds in the group $\mathbb{G}$, and that the hash function H is a random oracle and the underlying polynomial representation securely represents the intersection information, then the proposed pattern oriented private census scheme can guarantee that the only information the aggregator obtains is the targeted intersection of answers of all users or the k users who respond to the census.

Proof: Our proposed pattern oriented census scheme can be viewed as a combination of the underlying secret key distribution scheme and the intersection information extractor proposed by Lin *et al.* [13]. We guarantee the private census scheme satisfies the "encrypt-once security" condition by using different index i for each partial ciphertext. Therefore, we have $\forall u_j, \forall c_{ji}$, the tuple (u_j, i, c_{ji}) would not appear twice in the encryption query. This would guarantee the aggregator only obtains the summation of the polynomial coefficients as long as decisional DH assumption holds in the underlying group. By the identical security argument in the intersection extractor scheme, the aggregator only obtains the targeted intersection information as long as the polynomial representation technique is secure. Hence, we can conclude the proof if the underlying secret key distribution protocol is reduced to the decisional DH assumption. □

The above solution gives the aggregator little control over the threshold k since it totally depends on the number of users willing to provide the real answers. However, the aggregator might wish to exert more control on k. We propose using anonymous communication as a solution for this scenario. At the beginning of the system, upon the aggregator publishing his private pattern census, the individual user can submit a sign indicating his willingness to provide the real answer through the anonymous communication channel. The aggregator counts the number of users willing to answer correctly until it reaches his preset threshold k, and publishes a message stating that the number of users is enough and stops users from sending more positive sign. Then they run the modified pattern census just as introduced in the above paragraph. The user who submits a positive sign provides the real answer, while the others act as shadow users.

The pattern oriented private census scheme can be viewed as a filter for the subsequent census. The aggregator can roughly obtain an answer to the following conjunction question: "k/n

percent of users have property 1 **AND** property 2, etc." through the modified pattern oriented private census scheme. Using this information as an indication, the aggregator can further design a more specific census to investigate exactly how much percentage of the users have those properties simultaneously. We note that the question list would consist of much fewer questions since the aggregator can now only focus on investigating a small portion of intersection property of much more significant implication rather than probing blindly in all the intersection space which could be of exponential size.

VII. Conditional Probability Estimation and Association Rule Mining

The proposed pattern oriented private census scheme combined with the general census scheme can serve as a powerful tool for computing baseline inference such as conditional probability and a more sophisticated application such as association rule mining.

The threshold pattern oriented census scheme provides an aggregated answer such as "more than 30% users consume 100 W per hour in a certain region **AND** have more than annual income exceeding 100 000 dollars." The general census scheme proposed later can answer exactly what percentage of users have both the above two properties, or just one of the properties. All these answers can be pulled together to compute the conditional probability using the equation $P(A|B) = P(A \bigcap B)/P(B)$, which could be vital for some classical learning technique such as Bayesian learning.

The other application of the pattern oriented census scheme is association rule mining. The definition for association rule can be found in [15]:

Definition 1: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Let DB be a set of transactions, where each transaction T is an item set such that $T \subseteq I$. Given an item set $X \subseteq I$, a transaction T *contains* X iff $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$ where $X \subseteq I, Y \subseteq I$ and $X \bigcap Y = \emptyset$. The rule $X \Rightarrow Y$ has *support* s in the transaction database DB if s% of transactions in DB contain $X \bigcup Y$. The association rule holds in the transaction database DB with *confidence* c if c% of transactions in DB that contain X also contains Y.

Association rule mining aims to figure out all rules with support and confidence higher than certain predefined threshold support and confidence. The proposed private census schemes can be viewed as an application of multi-party private set intersection or union schemes, which is generally considered as a powerful tool for association rule mining [23]. Hence, our proposed threshold private pattern census combined with the general private census can be directly applied to construct a private association rule mining protocol. Assuming the confidence is c, then the aggregator can run a threshold private pattern census with threshold $k = c$ and thus find out all the X and Y in the question list such that $X \bigcap Y = \emptyset$ and the number of the users who choose the intersection of choices X **AND** Y is $\geq c$. Then, the aggregator can publish a general census to calculate the number of users who choose the choices $X \bigcup Y$. The aggregator will then count all the $X \bigcup Y$ with more than s number of users assuming s is the threshold for support.

VIII. Correctness of Metering Data

It is easy to observe that the answers or responses related to metering data play an important role in the profiling. It is necessary to provide a mechanism to ensure individual answers are indeed "correct" according to the returned metering device data. We propose to use the non-interactive zero knowledge proof combined with homomorphic commitment scheme and aggregated signature as the underlying tool for the verification on the correctness of the answers. Our scheme makes the same assumption as most of the smart metering based systems [1], [16] in the sense that metering devices are tamper-resistant.

Advanced metering devices are supposed to deliver metering data in each short period of time, such as half an hour. The question in the census related to metering data could either correspond to a single measurement period or K measurement periods. A metering device commits to metering data $d_i, i \in [1, K]$ for each metering time period t and generates a signature $\mathrm{Sig}(t \| \mathrm{Com}_i)$ for each commitment $\mathrm{Com}_i = g^{d_i} h^{r'}$ (We use Pedersen's commitment scheme in here [24]). A user collects all the commitments and the respective signatures, and generates the commitment for the summation $g^{\sum_{i=1}^{K} d_i} h^r$ using the additive homomorphic property of the commitment scheme and the aggregate signature $\mathrm{ASig}(\{t \| \mathrm{Com}_i\}_{i=1}^{K})$ derived from $\mathrm{Sig}(t \| \mathrm{Com}_i)$, where t indicates the generation time of the metering data, and $\mathrm{Sig}(t \| \mathrm{Com}_i)$ is the individual signature for each commitment.

The verification approach can be divided into two categories dependent on the underlying secret key distribution protocols: the semi-decentralized model and the fully decentralized model. The verification for the semi-decentralized model is straightforward and similar to that of the interactive model in [1] and thus is omitted. We mainly focus on the verification mechanism for the fully decentralized model.

For the general census scheme, u_j presents a non-interactive zero-knowledge proof (NIZK) as follows. $\mathrm{NIZK}\{(A_j(I) = \sum_{i=1}^{K} d_i, r, SK_j) : g^{\sum_{i=1}^{K} d_i} h^r = \mathrm{Com}(\sum_{i=1}^{K} d_i, r) \wedge g^{A_j(I)} H(I)^{SK_j} = g^{\sum_{i=1}^{K} d_i} H(I)^{SK_j}\}$, where $\mathrm{Com}(\sum_{i=1}^{K} d_i, r)$ is a commitment for $\sum_{i=1}^{K} d_i$ and the respective open value r, and I is the respective question index. The aggregator first verifies $\mathrm{ASig}(\{\mathrm{Com}_i\}_{i=1}^{K})$ corresponding to the metering device public key $PK_{(m)}$ and then checks whether $\mathrm{Com}(\sum_{i=1}^{K} d_i, r)$ is indeed the commitment for $\sum_{i=1}^{K} d_i$ and r using the homomorphic property of individual commitment Com_i, and finally checks the correctness of NIZK. The correctness of the submitted answer is verified when all the above verification does not fail.

As for the pattern oriented census scheme, the non-interactive zero-knowledge proof is modified as follows: $\mathrm{NIZK}\{(\sum_{i=1}^{K} d_i, r, SK_j, R_1(x), R_2(x), L_i \leq \sum_{i=1}^{K} d_i \leq R_i, A_j(I_1) = L_i, A_j(I_2) = R_i) : g^{\sum_{i=1}^{K} d_i} h^r = \mathrm{Com}(\sum_{i=1}^{K} d_i, r) \wedge g^{(x - A_j(I_1))R_1(x)} \prod_{i=0}^{4V-2V'} H(i)^{SK_j x^i} = \prod_{i=0}^{4V-2V'} C_{ji}^{x^i} \wedge g^{(x - A_j(I_2))R_2(x)} \prod_{i=0}^{4V-2V'} H(i)^{SK_j x^i} = \prod_{i=0}^{4V-2V'} C_{ji}^{x^i}\}$, where $\mathrm{Com}(\sum_{i=1}^{K} d_i, r)$ is still a commitment to $\sum_{i=1}^{K} d_i$ and the respective open value r. $R_1(x)$ and $R_2(x)$

are the randomization polynomials, respectively. The aggregator first verifies the aggregate signature and commitment in the same manner as in the general census scheme, and finally checks the correctness of NIZK, which will guarantee that the user indeed chooses the correct interval when he provides the real answer. Both the above zero-knowledge proof schemes can be efficiently realized by using the efficient non-interactive proof scheme proposed in [25]. The above verification steps can prevent the user from manipulating the metering data related answer. The answer for other properties such as the individual socio-economic background can still be verified by other mechanism such as anonymous credential, which might be more communication and computation intensive. However, we argue that there would be little incentive for users to fabricate metering data or other socio-economic background related data because the effect of individual data has negligible impact on the aggregated result considering the scale of the system.

IX. Performance Analysis

In this section, we evaluate our proposed schemes. The underlying secret key distribution protocol requires $O(n^2)$ messages to be exchanged when it corresponds to the fully decentralized model. Each individual user is required to perform $O(n)$ public key encryptions and decryptions. When it comes to the semi-decentralized model, the exchange messages are of size $O(n\ell)$. An individual user performs $O(\ell)$ public key encryptions, and the aggregator computes $O(n)$ public key decryptions. We note that the secret key distribution protocol only needs to be executed once during the system initialization and multiple instances of private census scheme can be run afterwards. Each individual user sends N group elements to the aggregator in the general private census scheme while is required to send $4V - 2V' + 1$ group elements to the aggregator in the pattern oriented census scheme. The general private census scheme requires $2N$ exponentiations and N module multiplications per user while the pattern oriented census scheme requires $8V - 4V' + 2$ exponentiations and $4V - 2V'$ multiplications. We note that we ignore the computational cost for polynomial multiplication in the exponent since it is negligible compared with the exponentiations and multiplications of group elements. The aggregator in the general private census scheme is required to accomplish $O(N(n\mathrm{Max} - \mathrm{Min}))$ exponentiations, which could be further reduced to $O(N\sqrt{n\mathrm{Max} - \mathrm{Min}})$ provided that the Pollard's lambda method is applied. In the pattern oriented census scheme, the aggregator needs to complete $(n - 1)(4V - 2V' + 1)$ group element multiplications and $(2V' + 2\sum_{i\in[V'+1,V]} G_i)(4V - 2V' + 1)$ exponentiations. It is noteworthy that the major computation workload for the aggregator introduced by the exponentiations does not depend on the number n of system users, but mostly depends on V and V', which are further determined by how the question list is formed.

We implement the two proposed private census schemes and the underlying secret key distribution protocol. We employ the real data from a recent investigation [26] on the relationship between energy consumption and temperature in Sydney, Australia to test the performance of the general private census scheme. There are totally 600 households involved in the system. It takes roughly 10 ms to accomplish a modular exponentiation of 1024 bit prime modular for a user with a 412 MHz smartphone. In the worst case of the proposed secret key distribution protocol, each user needs to perform 600 public key encryptions, which corresponds to roughly $1200 * 10\ \mathrm{ms} = 12$ s computations. Each user delivers two group elements to the other user, which is of size 2 KB. We consider the aggregation of hourly energy consumption data versus the temperature variation in the general private census scheme. The maximum hourly energy consumption for an individual appliance is 500 W and the minimum is about 50 W according to the data provided by the investigation. The maximum and minimum temperature is 10 °C and 25 °C, respectively. In order to obtain the average energy consumption for each temperature in $[10, 25]$, the aggregator needs to post a census consisting of 17 questions. Hence, to respond to the census, an individual user needs to generate 17 answers, which corresponds to $17 * 2$ modular exponentiation computations, which takes about $34 * 10\ \mathrm{ms} = 0.34$ s. The data generated by the individual user is of size $34 * 1024\ \mathrm{bits} = 34$ KB. The aggregator will need to perform $n * \mathrm{Max} - \mathrm{Min} = 600 * 500 - 50 = 299\,950$ times of modular exponentiations to compute the average energy consumption for all the possible temperature values in $[10, 25]$. The aggregator can use a desktop to compute the average energy consumption, and it takes roughly 0.3 ms to compute a modular exponentiation in a modern 64-bit desktop [18]. Therefore, it would take roughly $299950 * 0.3\ \mathrm{ms} = 90$ s at most to compute the answer for each temperature value. It is noteworthy that the individual computation cost remains constant even if the system users increases while the communication overhead does grow with the number of system users.

For the proposed pattern oriented census scheme, we use the real data from another recent investigation [27] on the relationship between the state (turn on or off) of individual appliances at home and the energy consumption at a particular time interval of a day. There are five time intervals and six appliances according to the investigation. Therefore, there will be six binary questions and five interval questions in order to calculate the correlation shown in Table I in [27], and hence $V' = 6$ and $V = 11$. The grain levels for the five time intervals are 10, 30, 15, 30, 30 assuming each interval choice is $[L_j, L_j + 1]$W in the investigation. Therefore, each user needs to compute $8V - 4V' + 2$ modular exponentiations, which takes roughly $(8 * 11 - 4 * 6 + 2) * 10\ \mathrm{ms} = 0.66$ s and the individual ciphertext size is $(4 * 11 - 2 * 6 + 1) * 1024/2\ \mathrm{bits} = 16.5$ KB. It would take the aggregator roughly $(2V' + 2\sum_{i\in[V'+1,V]} G_i)(4V - 2V' + 1) * 0.3\ \mathrm{ms} = (2 * 6 + 2(10 + 30 + 15 + 30 + 30))(4 * 11 - 2 * 6 + 1) * 0.3/1000\ \mathrm{s} = 2.3958$ s to calculate the intersection answer. The individual computing time and communication overhead dependent on V and V' can be found in Fig. 2, and the aggregator computing time dependent on V and V' can also be found in Fig. 3. However, the trivial approach can only reveal the answer which is chosen by all the users. In order to calculate the intersection answer chosen by k users, the individual user needs to compute $(4V' + 4\sum_{i=V'+1}^{V} G_i + 1) * 2$ modular exponentiations at most, which takes roughly $(4 * 6 + 4(10 + 30 + 15 + 30 + 30) + 1) * 2 * 10\ \mathrm{ms} = 9.7$ s and the ciphertext size is $(4 * 6 + 4(10 + 30 + 15 + 30 + 30) + 1) * 1024 = 485$ KB. The computation time for the aggregator in this case is identical to the case when $k = n$, i.e., roughly, 2.4 s.

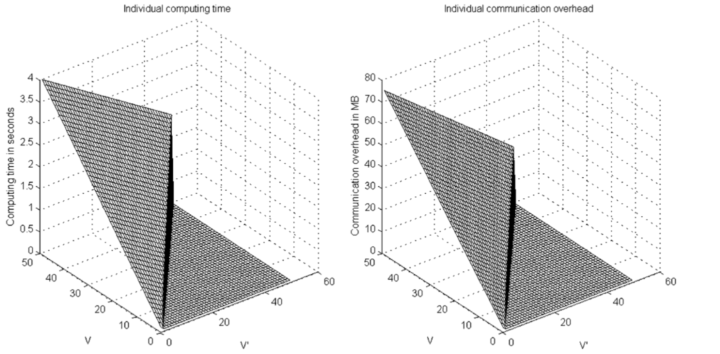

Fig. 2. The individual computing time and communication overhead.

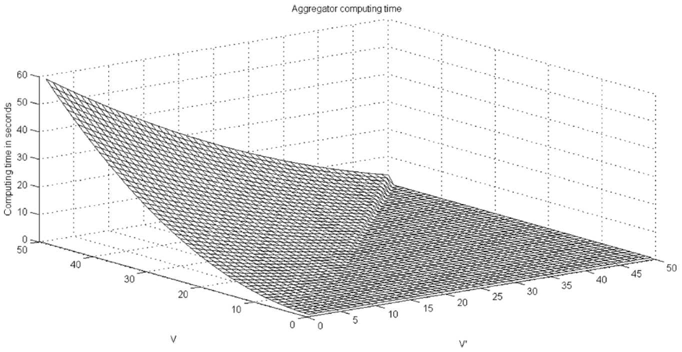

Fig. 3. The aggregator computing time.

X. Conclusion

This paper proposes two schemes to support privacy-aware and efficient dynamic profiling and statistical data extraction for smart sustainable energy systems. The first scheme enables an aggregator to efficiently extract the summation information from the submitted individual responses. The second pattern oriented privacy-aware profiling scheme can further enable the aggregator to efficiently extract the intersection information, which can answer the conjunction query. We also demonstrate how to use the latter scheme to perform private baseline inference and mining association rule through combining with the first scheme. Since associate rule mining [14] can discover interesting relationships between variables in large databases [15], the proposed schemes are believed to be beneficial to the smart grid energy policy and marketing strategy design in the long run.

References

[1] K. Kursawe, G. Danezis, and M. Kohlweiss, "Privacy-friendly aggregation for the smart-grid," in *Proc. PETS*, 2011, pp. 175–191.

[2] Y. Yohanis, J. Mondol, A. Wright, and B. Norton, "Real-life energy use in the uk: How occupancy and dwelling characteristics affect domestic electricity use," *Energy Buildings*, vol. 40, no. 6, pp. 1053–1059, 2008.

[3] R. A. Popa, A. J. Blumberg, H. Balakrishnan, and F. H. Li, "Privacy and accountability for location-based aggregate statistics," in *Proc. ACM Conf. Comput. Commun. Security*, 2011, pp. 653–666.

[4] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," *Proc. Ubicomp*, pp. 89–98, 2011.

[5] K. Budka, J. Deshpande, J. Hobby, Y.-J. Kim, V. Kolesnikov, W. Lee, T. Reddington, M. Thottan, C. A. White, J.-I. Choi, J. Hong, J. Kim, W. Ko, Y.-W. Nam, and S.-Y. Sohn, "Geri-bell labs smart grid research focus: Economic modeling, networking, and security & privacy," *Proc. IEEE SmartGridComm*, 2010.

[6] A. Markandya, I. Galarraga, and Gonzalez-Eguino*, Handbook for Sustainable Energy*. Cheltenham, U.K.: Edward Elgar Publ., 2011.

[7] F. D. Garcia *et al.*, J. Cuellar, Ed., "Privacy-friendly energy-metering via homomorphic encryption," in *Proc. 6th Workshop Security Trust Manage. (STM 2010)*, 2011, vol. 6710, pp. 226–238.

[8] K. Herter, P. McAuliffec, and A. Rosenfeld, "An exploratory analysis of california residential customer response to critical peak pricing of electricity," *Energy*, pp. 25–34, 2007.

[9] J. Snook, "Driving sustainable behavior in the mainstream consumer: Leveraging behavioral economics to minimize household energy consumption," Ph.D. dissertation, Duke Univ., Durham, NJ, 2011.

[10] I. H. Rowlands and I. M. Furst, "The cost impacts of a mandatory move to time-of-use pricing on residential customers: An Ontario (Canada) case-study," *Energy Efficiency*, vol. 4, pp. 571–585, 2011.

[11] S. Darby, "Social learning and public policy lessons from an energy-conscious village," *Energy Policy*, vol. 34, no. 17, pp. 2929–2940, 2006.

[12] M. Enev, S. Gupta, T. Kohno, S. N. Patel, and S. N. Patel, "Televisions, video privacy, and powerline electromagnetic interference," in *Proc. ACM Conf. Comput. Commun. Security*, 2011, pp. 537–550.

[13] H. Lin, Y. Fang, and Z. Cao, "Private information extraction over online social networks," *IACR Cryptol. Print Archive,* 2011 [Online]. Available: http://eprint.iacr.org/2011/446

[14] R. Agrawal, T. Imielinski, and A. N. Swami, P. Buneman and S. Jajodia, Eds., "Mining association rules between sets of items in large databases," in *Proc. 1993 ACM SIGMOD Int. Conf. Manage. Data*, Washington, DC, May. 26–28, 1993, pp. 207–216.

[15] R. Agrawal and R. Srikant, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds., "Fast algorithms for mining association rules in large databases," in *, Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, Santiago de Chile, Chile, Sep. 12–15, 1994, pp. 487–499.

[16] A. Molina-Markham, G. Danezis, K. Fu, P. J. Shenoy, and D. E. Irwin, "Designing privacy-preserving smart meters with lowcost microcontrollers," *IACR Cryptology ePrint Archive*, vol. 2011, p. 544, 2011.

[17] G. Danezis, M. Kohlweiss, and A. Rial, "Differentially private billing with rebates," *Inf. Hiding*, pp. 148–162, 2011.

[18] E. Shi, T.-H. H. Chan, E. G. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Proc. NDSS Symp.*, 2011.

[19] J. Balasch, A. Rial, C. Troncoso, B. Preneel, I. Verbauwhede, and C. Geuens, "Pretp: Privacy-preserving electronic toll pricing," in *Proc. USENIX Security Symp.*, 2010, pp. 63–78.

[20] A. Rial and G. Danezis, "Privacy-preserving smart metering," in *Proc. WPES*, 2011, pp. 49–60.

[21] R. Gennaro, S. Jarecki, H. Krawczyk, and T. Rabin, "Robust threshold dss signatures," *Inf. Comput.*, vol. 164, no. 1, pp. 54–84, 2001.

[22] L. Kissner and D. X. Song, "Privacy-preserving set operations," in *Proc. CRYPTO*, 2005, pp. 241–257.

[23] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Zhu, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explorations*, vol. 4, no. 2, 2003.

[24] T. Pedersen, "Non-interactive and information-theoretic secure verifiable secret sharing," in *Proc. Adv. Cryptol.—CRYPTO91*, 1992, pp. 129–140.

[25] J. Groth and A. Sahai, "Efficient non-interactive proof systems for bilinear groups," *Proc. Electron. Colloq. Comput. Complexity (ECCC)*, vol. 14, no. 053, 2007.

[26] M. Hart and R. de Dear, "Weather sensitivity in household appliance energy end-use," *Energy Building*, vol. 36, pp. 161–174, 2004.

[27] A. Jenny, J. R. D. Lopez, and H.-J. Mosler, "Household energy use patterns and social organisation for optimal energy management in a multi-user solar energy system," *Progr. Photovoltaics: Res. Appl.*, vol. 14, pp. 353–362, 2006.

Authors' photographs and biographies not available at the time of publication.