Seer Grid: Privacy and Utility Implications of Two-Level Load Prediction in Smart Grids

Arash Boustani, *Student Member, IEEE,* Anindya Maiti, *Student Member, IEEE,*
Sina Yousefian Jazi, Murtuza Jadliwala, *Member, IEEE,* and Vinod Namboodiri, *Senior Member, IEEE*

Abstract—We propose "Seer Grid", a novel two-level energy consumption prediction framework for smart grids, aimed to decrease the trade-off between privacy requirements (of the customer) and data utility requirements (of the energy company (EC)). The first-level prediction at the household level is performed by each smart meter (SM), and the predicted energy consumption pattern (instead of the actual energy usage data) is reported to a cluster head (CH). Then, a second-level prediction at the neighborhood level is done by the CH which predicts the energy spikes in the neighborhood or cluster and shares it with the EC. Our two-level prediction mechanism is designed such that it preserves the correlation between the predicted and actual energy consumption patterns at the cluster level and removes this correlation in the predicted data communicated by each SM to the CH. This maintains the usefulness of the cluster-level energy consumption data communicated to the EC, while preserving the privacy of the household-level energy consumption data against the CH (and thus the EC). Our evaluation results show that Seer Grid is successful in hiding private consumption patterns at the household-level while still being able to accurately predict energy consumption at the neighborhood-level.

Index Terms—Smart grid, smart meter, load prediction, privacy, data utility.

1 Introduction

As part of the future smart electricity grid initiative, a smart grid communication network (SGN) is a large-scale integration of information and communication technologies within the electricity generation, transmission, and distribution systems of the traditional electricity grid. A combination of various *smart* technologies at different levels of the SGN promotes efficiency, reliability and stability in operations of the smart grid. One indispensable piece of technology in a SGN is a smart meter (SM) which collects and periodically reports the energy usage or consumption information of the customers to the electric (a.k.a. utility) company (EC), which in turn facilitates highly efficient energy generation and distribution and helps the EC to cope with changes in energy demand and supply. The monetary and natural resource savings due to the improved efficiency is a major factor in the fast growing adoption of SMs, with predictions that about 800 million SMs will be in use globally by 2020 [1]. Despite their tremendous importance in a SGN, SMs can also be easily exploited by malicious adversaries (including the EC) who may attempt to infer private customer information from reported energy consumption patterns, such as occupancy of the house [2], specific appliances being used [3], and even daily routine of the residents [4] [5].

Various techniques for overcoming privacy issues due to the energy usage information generated and shared by SMs have been proposed in the research literature, and these solutions have primarily followed one of the following two approaches: (i) completely obscure the individual SM data from the perceived adversary, or (ii) hide privacy-sensitive signatures or patterns from the individual SM data by perturbation or down-sampling. In the first direction, protocols that take advantage of the homomorphic properties of public-key cryptographic algorithms to perform neighborhood-level aggregation of SM data have been proposed in the literature [6] [7]. These protocols enable the EC to learn the actual aggregated energy consumption information (at a neighborhood level) without leaking individual customer-specific information to the aggregator. In the second direction, many approaches have been proposed to efficiently perturb energy consumption data in order to meet certain privacy requirements. In-residence storage batteries have been employed to flatten or mask variances in the load or electricity usage information [8] [9]. Similarly, controlled perturbation [10] [11] [12] and down-sampling [13] [14] of the energy consumption data to mask specific load signatures of appliances have also been attempted.

However, as pointed out by [12] and [13], the degree of correlation between the actual energy consumption and the data output by a privacy-preserving technique typically characterizes a trade-off between privacy and utility (or usefulness). Higher correlation with the actual ground-truth makes the perturbed data more useful but reveals private information, whereas lower correlation (or increased perturbation) is good for privacy but reduces data usefulness or utility. As protocols following the first approach do not really perturb the electricity consumption data, the utility of the data (or any function computed from the data) is high. Also, as this data is cryptographically obscured from the aggregator, there is no leakage of private customer information. However, protocols using public-key cryptography are non-trivial to implement in practice and have very high computation and communication overhead [15]. Perturbation mechanisms, such as the ones using storage batteries

• *A. Boustani, A. Maiti, S. Y. Jazi, M. Jadliwala, and V. Namboodiri are with the Wichita State University, Wichita, Kansas 67260, USA.*
Email: {axboustani, axmaiti, sxyousefianjazi, murtuza.jadliwala, vinod.namboodiri}@wichita.edu

Manuscript received December 2, 2015; revised May 1, 2016.

[8] [9], are effective in masking private usage patterns, but installing and maintaining large capacity batteries in every household is shown to be economically non-viable [16]. Similarly, Dong et al. [13] show that performance of smart grid operations can degrade due to reduction in sampling frequency. Other perturbation mechanisms, such as [12], that attempt to strike a good balance between privacy and data-utility by masking or suppressing specific appliance signatures assume that individual appliance electricity consumption information is readily available (or can be easily separated from the overall data) which may not always be feasible. Given the above state-of-the-art, we feel that both data hiding and data perturbation approaches have inherent limitations, which motivates us to explore alternate paradigms (beyond hiding and perturbation).

Our goal in this work is to explore alternate practical designs for privacy-sensitive generation and sharing of energy consumption information from the SMs to the EC which enables effective operation of the EC in terms of accurately predicting future demand and electricity generation and distribution. In order to achieve this goal, we move away from the classical perturbation/data-hiding techniques and use learning-based prediction mechanisms to generate (or predict) energy consumption patterns shared by SMs. Our prediction mechanism will replace variances in the individual household-level actual energy consumption patterns (which is typically indicative of loads) with relatively smoother patterns that are free of load signatures but accurate enough to be useful in predicting energy consumption at the neighborhood level (which is the one that is actually used by the EC). Due to this, privacy-sensitive inference attacks will be much harder on the household-level data shared by the SM without significantly impacting the demand-response and electricity generation/distribution calculations at the EC.

With Seer Grid[1], we propose a household-level prediction scheme comprising of a statistical learning algorithm (trained using past consumption pattern of the household) which predicts an entire day's electricity consumption pattern one day in advance. This prediction can be performed locally on the SM, on a local energy management unit or on a computing device that connects to such a unit. The household electricity consumption pattern predicted locally at the SM, with the load or appliance signatures masked or flattened, is then reported to an aggregator or data concentrator (referred here as a cluster head or CH) at the beginning of each day. All SMs within a neighborhood or cluster report their energy consumption predictions to their respective CH who in turn forwards an aggregated prediction (as described below) to the EC. As our localized prediction flattens or eliminates sharp variations (which may indicate load signatures) in the predicted consumption at the SM or household level, this difference can add up significantly while aggregating predictions for multiple households in a neighborhood or a cluster. This can reduce the accuracy of the aggregated prediction, thereby adversely impacting its utility or usefulness to the EC. In order to restore this utility lost due to prediction at the SM level, we introduce a second level of energy load prediction at the CH for compensating the difference in the aggregate of predicted and actual energy usage of individual SMs in the cluster. CH performs the spike prediction based on past energy consumption pattern of the entire neighborhood or cluster, and then reports the result of the second level prediction, in addition to summation of first level predictions, to EC just before beginning of each day. EC can then use this cluster or neighborhood-wide load prediction to efficiently control electricity generation and distribution. To ensure fail-proof operation of the SGN in case of major prediction errors, we also incorporate real-time and privacy-preserving reporting of the aggregated variance between actual and predicted energy consumption of all SMs in the cluster.

Seer Grid's two-level prediction mechanism offers several advantages over traditional privacy-preserving energy data reporting schemes in the literature. Unlike data hiding schemes that require several encryption operations at the SM or household level per day (once every reporting interval), our prediction and reporting operation is performed just once per day. Moreover, Seer Grid is communication-efficient (as no additional data or overhead needs to be communicated), does not require any specialized hardware (e.g., storage batteries) and does not need access to appliance-level consumption patterns. In rest of the paper, we first describe the generic SGN architecture and capabilities of the assumed adversary in Section 3. In Section 4, we discuss the details of the proposed Seer Grid architecture and it's operation. In Section 5, we evaluate the Seer Grid architecture by performing extensive experimental simulations using real smart meter data. We empirically measure the correlation between predicted and actual consumption patterns at each level, using standardized metrics, to support our proposition of a practical SGN architecture which improves both privacy and utility of SM data. Comparative evaluation shows that Seer Grid's two-level prediction provides better privacy and utility, compared to just SM level perturbation techniques.

2 Related Work

Multiple schemes have been proposed for short term [17] [18] [19] and long term [20] load prediction at cluster level. Sevlian and Rajagopal [21] proposed short term electricity load forecasting on varying levels of aggregation, and concluded that aggregating more customers improves the relative forecasting performance only up to a specific point. Recently, smart meter based short-term load forecasting was proposed [22] [23], as a household's historic energy consumption pattern is a better predictor of peak load than any other observable variables. In contrast, Seer Grid uses two level of prediction to retain the privacy benefits of aggregation, and utility benefits of individual household prediction.

There have also been extensive research efforts that attempt to address privacy issues related to SM data release. Li et al. [6] proposed using Paillier cryptosystem's homomorphic property for distributed energy consumption data aggregation from SMs, where the EC is able to know only the aggregated data upon decryption of the aggregated cipher. Garcia and Jacob [7] proposed the combination of

1. According to Oxford Dictionary, *seer* is "a person who is supposed to be able, through supernatural insight, to see what the future holds." Through two-level energy prediction we enable insight into future demands, while simultaneously promoting consumer privacy.

additive secret sharing algorithms and cryptosystems with homomorphic property, in order to compute the aggregated energy consumption of a given set of users (for example, in a cluster) in a privacy preserving fashion. However, cryptosystems with homomorphic properties induce a large computational overhead on the SMs, and real-time reporting in short time interval is impractical [15]. Alternatively, McLaughlin et al. [9] proposed a non-intrusive load leveling model by using large capacity batteries. Large capacity batteries smoothen the energy consumption pattern and effectively help in hiding load signatures contained in actual consumption pattern. However, large batteries are economically inconvenient [16] due to their high capital cost and low energy-efficiency.

Privacy through anonymization tries to unlink the energy usage data from individual SMs [24]. However, anonymization may turn out to be ineffective, as Jawurek et al. [25] and Faisal et al. [26] demonstrated the feasibility of using household anomaly detection and behavior pattern to de-anonymize SM data. With limited computational capabilities and practicality in mind, researchers suggested the use of perturbation techniques for hiding load signatures. Consumer privacy can be preserved by deliberately introducing error into the energy usage data [10] [11] [27] [28], and such perturbation techniques often try to achieve differential privacy in order to reduce the privacy-utility trade-off [12]. However, the privacy-utility trade-off of SM data perturbation techniques can still be significant [29], which may not be admissible to ECs. In this paper, the proposed Seer Grid architecture aims to decrease the privacy-utility trade-off by using a two-level prediction scheme.

3 The Traditional SGN Architecture

We base our work on one of the most popular SGN architecture consisting of a three-level hierarchical network (Figure 1). At the lower level are the SMs, physically located in households of end-users or customers. At the middle level, each neighborhood has a CH, and SMs report energy consumptions to CH. Situated at the higher level is the EC, to which all CHs report aggregated load of their respective neighborhood. The load reporting from all CHs aids EC in optimizing generation and distribution of electricity. In real-life implementation, CH may be owned and operated by a third party or by the EC itself.

We assume a passive adversary who may try to infer personal information of customers based on accessible energy consumption data. Motivations can vary widely, such as financial gain from advertising agencies, health insurance companies trying to find unhealthy lifestyle of insurees, etc. If given access to actual energy consumption data, the adversary is computationally capable of carrying out inference attacks by analyzing the data. We also assume that the adversary can access energy consumption data reported to the CH and/or to the EC. However, CH and EC must be honest and cooperative for the protocol to function properly. Thus, CH and EC can be considered as honest but curious adversaries. We also consider any eavesdropper (eavesdropping communication between SM and CH, or CH and EC) as an adversary. All SMs are assumed to honestly and correctly follow the proposed protocol. As a result, we do not consider collusion attacks between SMs and CH, or between SMs and EC.

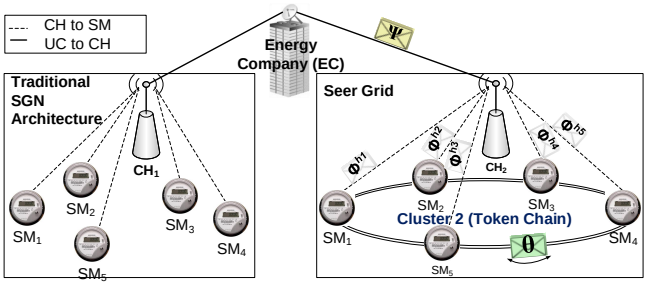


Fig. 1. Traditional SGN architecture on the left, and our proposed SGN architecture (details in Section 4) on the right.

4 Seer Grid

The primary distinction between the traditional SGN and Seer Grid is that, in Seer Grid SMs never report their actual energy consumption data; they report predicted energy consumption pattern instead. Similar to the traditional SGN architecture, Seer Grid also consists of a three level hierarchical network (Figure 1). At the lower level are the SMs, physically located in households. At the middle level, each neighborhood has a CH, and SMs report next day's predicted energy consumption patterns to CH. A second level prediction is performed by CH on the aggregated predicted patterns reported by all SMs belonging to the cluster. At the higher level is the EC, to which all CHs report the second level predicted energy forecast for their respective neighborhood. The predicted forecast from all CHs aids EC in optimizing generation and distribution of electricity. We assume that the CH is capable of measuring the actual electricity usage of the whole cluster for a given time interval[2], which is used to the form statistics used in the cluster level prediction. This is a reasonable assumption because all cluster level energy forecasting schemes [17] [18] [19] [20] rely on readings from a cluster level electricity meter. We also consider billing once as a month event, which can be done independently.

We carefully analyzed various statistical learning algorithms for predicting energy consumption patterns, in order to identify the algorithm apposite for preserving only the desired characteristics of the consumption pattern data. We first detail the constituents and properties of the consumption pattern data, followed by a discussion on how we select prediction algorithms for SM and CH. Readers should note that we use the following prediction algorithms as an example, in order to demonstrate the benefits of Seer Grid. Other suitable prediction algorithms can be used as well.

4.1 Prediction at SM Level

In a traditional SGN, SMs report energy usage data in short time intervals, where each report conveys the energy

2. CH is assumed to be equipped with a cluster level meter which measures energy being withdrawn by the entire cluster. When CH measures the electricity usage of the entire cluster, it does not violate privacy of individual household because of aggregation.

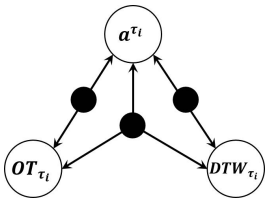


Fig. 2. Interaction between $a^{(\tau_i)}$ and OT_{τ_i} is 2-way. Interaction between $a^{(\tau_i)}$ and DTW_{τ_i} is also 2-way. And there exists a 3-way interaction between $a^{(\tau_i)}$, OT_{τ_i} and DTW_{τ_i}. The prediction model must learn these interactions in order to make effective predictions.

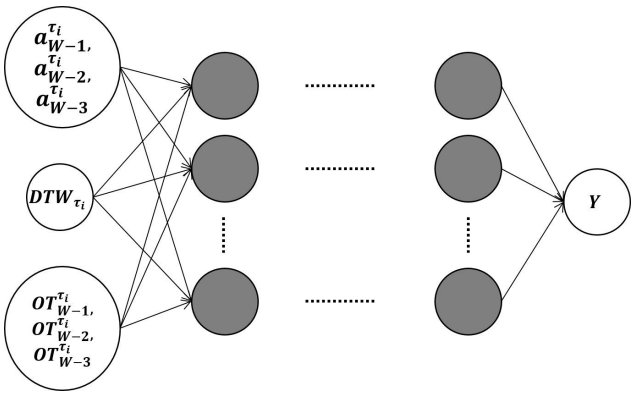


Fig. 3. The abstract structure of the MLP used of learning and prediction. $a^{(\tau_i)}_{W-1}$, $a^{(\tau_i)}_{W-2}$ and $a^{(\tau_i)}_{W-3}$ is the power usage in the τ_ith interval from last 3 weeks; DTW_{τ_i} represents day and time of the week,; and $OT^{(\tau_i)}_{W-1}$, $OT^{(\tau_i)}_{W-2}$ and $OT^{(\tau_i)}_{W-3}$ are the outdoor temperatures (in Fahrenheit) in the τ_ith interval from last 3 weeks.

consumed since the last reporting. Let us denote the actual daily SM energy consumption pattern of a household h_k as $\mathcal{A}^{h_k} = \{a^{(\tau_1)}, a^{(\tau_2)}, \cdots, a^{(\tau_n)}\}$, where $a^{(\tau_i)}$ is the energy consumed since $a^{(\tau_{i-1})}$. In Seer Grid, the goal of using a prediction model at the SM is to predict a pattern $\Phi^{h_k}_{day_j} = \{p^{(\tau_1)}, p^{(\tau_2)}, \cdots, p^{(\tau_n)}\}$, such that there is a high overlap between $\Phi^{h_k}_{day_j}$ and $\mathcal{A}^{h_k}_{day_j}$, but $\Phi^{h_k}_{day_j}$ is free of specific load signatures (such as spikes and plateaus). Predictive modeling leverages statistics to predict outcomes, i.e., the forecast of a day's consumption pattern is based on collection of past $\mathcal{A}^{h_k}$ (let's say for m days). After analyzing various factors that affect consumers' energy usage, we identified the input variables critical to the outcome of the prediction model as [i] power usage history in each time interval ($a^{(\tau_i)}$), [ii] outdoor temperature [3] in each interval (OT_{τ_i}), and [iii] day and time of the week (DTW_{τ_i}). Each day of the week is considered differently so as to improve prediction based on weekly routines [30]. All interactions present between these three variables is represented in Figure 2.

Classical time series forecasting techniques [31] use a statistical model for predicting future values based on previously observed values. However, such basic time series forecasting does not capture the complex interactions between different input variables, thus resulting in inferior

3. Many older SMs are not equipped with temperature sensors. In such cases, appropriate outdoor temperature values can be provided by the CH or EC.

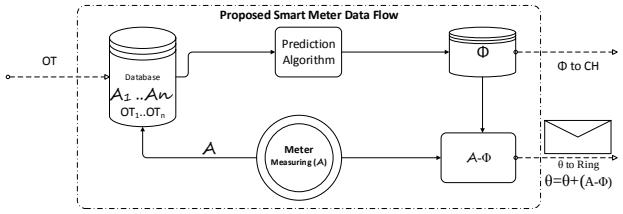


Fig. 4. Proposed SM data flow.

forecasting. Due to the highly complex interactions and some dependencies between input variables, multi-class classification and regression analysis will also result in non-optimal prediction. To achieve better prediction results, we employ structured prediction using supervised machine learning techniques. Among candidate machine learning techniques for structured prediction, we decided to use multi-layered perceptron (MLP) because it is specifically designed to discover the complex interactions among input variables. MLP is a feed-forward artificial neural network (ANN) model that uses a nonlinear activation function to map sets of input data onto a set of appropriate outputs. MLPs consisting of three or more layers (input, output, and one or more hidden layers) is called a deep neural network, where each node in one layer connects with a certain weight w_{pq} to every node in the following layer. The error in output of a node q in the nth training data point is represented as $e_q(n) = d_q(n) - y_q(n)$, where d is the target value and y is the value produced by the perceptron. The calculated error for each training data point is used to make corrections to the weights of the node as $\mathcal{E}(n) = \frac{1}{2}\sum_q e^2_q(n)$, which in turn minimizes the error in the entire output of the ANN. Change in each weight during an epoch is calculated as $\Delta w_{qp}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_q(n)} y_p(n)$, where y_p is the output of the previous neuron and η is the learning rate.

In the learning phase of our MLP execution, for each epoch we input power usage history of past three weeks recorded in 5 minute intervals. Outdoor temperature for the corresponding interval and day of the week is also fed in each epoch (Figure 3). The output of the ANN is a structured object (Y) containing multiple possible $\Phi^{h_k}_{day_j}$ for next day. Given that the next day's temperature forecast and day of the week is known, the structure object is parsed for the matching $\Phi^{h_k}_{day_j}$. More details about the MLP specifications used in our simulation experiments can be found in Section 5. Additionally, Connor et al. [32] demonstrated that neural networks trained on filtered data can perform better predictions than neural networks trained on unfiltered time series. Therefore, we use a low-pass filter over SM training data which shaves any load pattern for energy consumption above 4kW. The value 4kW was empirically determined based on the observation that in our training data, more than 94% of data samples are lower than 4kW.

A distributed model of SMs is used in our proposed SGN model, where the first level prediction is performed independently on all SMs belonging to the SGN. The prediction algorithm running inside the SM of a household h_k locally stores a small database (Figure 4), containing actual consumption patterns $\mathcal{A}^{h_k}$ and outdoor temperature

measurements OT_i from last m days. Each day, the $\mathcal{A}^{h_k}$ and OT_i values are used to train the MLP and predict the $\Phi^{h_k}_{day_j}$ for next (j-th) day. Also, at the end of a day, that day's actual consumption pattern $\mathcal{A}^{h_k}_{day_{j-1}}$ is inserted into the queue of the database and the oldest actual consumption pattern $\mathcal{A}^{h_k}_{day_{j-m+1}}$ is removed. As mentioned before, $\Phi^{h_k}_{day_j}$ is computed and reported only once (before beginning of) each day. All communications between SM and CH for reporting $\Phi^{h_k}_{day_j}$ are assumed to be symmetrically encrypted, for example, by using AES [33].

4.2 Prediction at CH Level

The purpose of using prediction at SM is to remove specific load signatures (such as spikes and plateaus) from $\mathcal{A}^{h_k}_{day_j}$. Although the missing spikes and plateaus from the SM of one household represent a minuscule amount of energy for the grid, spikes and plateaus from multiple households in a cluster can add up to a significant amount of unpredicted energy. This accumulated error in prediction can affect processes that would use the predicted data, for example, intelligent electricity distribution, demand-response, etc. Therefore, we introduce another level of statistical prediction at the CH based on historical load profile of the cluster, while also factoring in individual predictions from all SMs in the cluster $\{\Phi^{h_1}_{day_j}, \Phi^{h_2}_{day_j}, \Phi^{h_3}_{day_j}, ...\}$.

As mentioned earlier, the CH is assumed to have load measurement capability to measure total energy consumption in it's neighborhood. The meter measures the energy injected into the entire cluster, without having access to actual individual SM readings at any point. As a result, CH can easily calculate the difference between the aggregated predicted values which are gathered from SMs and the measure of actual total energy consumption in the cluster. The proposed algorithm (Algorithm 1) uses average of difference between past load predictions and actual loads of the entire cluster ($\Lambda_{day_d} = \{\lambda^{(\tau_1)}, \lambda^{(\tau_2)}, \cdots, \lambda^{(\tau_n)}\}$), in order to complement missing loads. The output of the algorithm $\Psi_{day_j} = \{\psi^{(\tau_1)}, \psi^{(\tau_2)}, \cdots, \psi^{(\tau_n)}\}$ is the prediction for the whole cluster reported to CH, where $\psi^{(\tau_i)} = \delta^{(\tau_i)} + \sum_k p^{(\tau_i)}$ and $\delta^{(\tau_i)} = \frac{\sum_{d=j-m}^{d=j-1}\{\lambda^{(\tau_i)}_{day_d} - \sum_k p^{(\tau_i)}_{day_d}\}}{m}$.

Algorithm 1 Prediction algorithm executed by CH.

Prediction Function (for day j)

Define new $\Psi_{day_j} = \{\psi^{(\tau_1)}, \psi^{(\tau_2)}, \cdots, \psi^{(\tau_{288})}\}$

for $k = 1$ to K (K households in the cluster) **do**
  $\sum p^{(\tau_i)}_{day_j}$
end for

for $i = 1$ to 288 (5 minutes time intervals for 24 hours) **do**
  for $d = 1$ to m (m days to historical data) **do**
    $\delta^{(\tau_i)} = \lambda^{(\tau_i)}_{day_d} - \sum_k p^{(\tau_i)}_{day_d}$
  end for
  $\delta^{(\tau_i)} = \frac{\delta^{(\tau_i)}}{m}$
  $\psi^{(\tau_i)} = \delta^{(\tau_i)} + \sum_k p^{(\tau_i)}$
end for

Report Ψ_{day_j} to CH

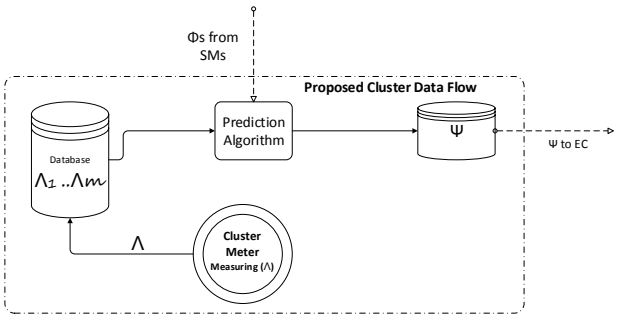


Fig. 5. Proposed CH data flow.

For the second level prediction, the CH accumulates all $\Phi^{h_k}_{day_j}$ in the cluster, adds the calculated $\delta^{(\tau_i)}$ to $\sum_k p^{(\tau)}$ for each time interval (τ), and reports the resulted pattern Ψ_{day_j} to the EC. CH also stores a database of past Λ and $\sum_k p^{(\tau)}$ values from last m days, which is updated at the end of each day (Figure 5).

4.3 Privacy Preserving Real-time Monitoring

The predicted pattern Ψ_{day_j} is a refined estimate of next day's energy consumption at the cluster level, as compared to individual SM predictions $\Phi^{h_k}_{day_j}$. However, there may occur unexpected events which are not captured by the input variables of our prediction system, for example severe weather conditions, natural disasters, etc. To ensure proper functioning of SGN in case of an unexpected power demand, we incorporate a real-time reporting system in our architecture to measure the difference in actual and predicted energy consumption of all households. But, directly reporting difference in actual and predicted energy consumption pattern to CH defeats our goal of privacy, because CH can add back the difference to predicted pattern to obtain the actual pattern of individual SMs. So, the real-time reporting system uses a "token chain" mechanism to aggregate the difference in actual and predicted energy consumption pattern for all SMs in the cluster. The token chain design design can be based on existing energy-efficient token passing mechanisms designed for ad-hoc wireless sensors networks [34] [35] and smart grid networks [36]. In the logical chain of all SMs belonging to a cluster, a token is circulated across all SMs (as shown in Figure 1) for aggregation of difference in actual and predicted energy consumption of the cluster. The difference in aggregated actual and predicted energy consumption $\theta^{(\tau)} = \sum_k (p^{(\tau)} - a^{(\tau)})$ in each time interval τ, can be used to handle unexpected demand events in real-time. Due to aggregation of the difference in actual and predicted energy consumption, individual household privacy is not compromised. Figure 4 illustrates how each SM adds their difference in actual and predicted energy consumption to the token. The final token value containing the aggregated difference in actual and predicted energy consumption of the cluster is reported to EC (via CH) for regulating generation and distribution, if necessary. To protect the token chain against eavesdropping inference attacks, all SMs symmetrically encrypt (and decrypt for addition)

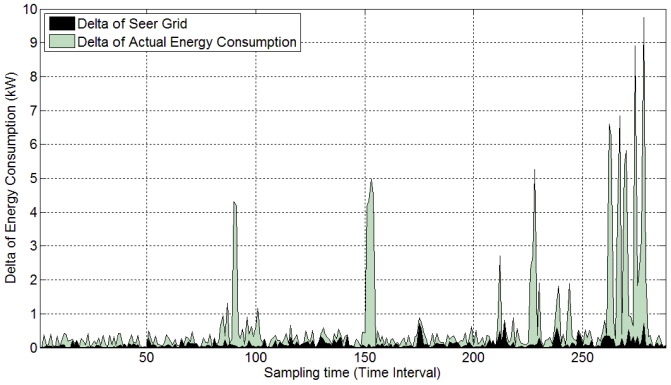


Fig. 6. Comparison of $\mathbb{D}p$ and $\mathbb{D}q$ over a test day. The lower values of $\mathbb{D}p$ means $\mathbb{P}$ is relatively "smoother".

the token using a shared secret, obtained by using a key exchange protocol, such the Diffie-Hellman protocol [37].

5 Evaluation

In order to validate the benefits of our proposed Seer Grid architecture, we conduct extensive experimental simulations using real smart meter data. In this section we present our experimental setup followed by results.

5.1 Experimental Setup

We use real SM data collected from residences equipped with $BS\ \ EN62053-21002003$ smart meters. The data was recorded in East Midlands, UK in the year 2008 [38]. The fabricated cluster we consider for evaluation consists of 5 households, each having one smart meter. Envisioning the limited memory that SMs may have, we limit the use of historical data in our experiments to three weeks, i.e, $m = 21$. Longer training period not only takes more storage space, but also makes less significant contribution in the prediction because of changing conditions (such as temperature) throughout the year. The ANN prediction algorithm is trained with data from past 21 days to predict the energy consumption for a test day. Every day, the last day's energy consumption information is added to the training set, and the oldest (22^{nd}) day's energy consumption information is removed from the training set. This helps account for changing seasons, and at the same time, limits memory requirements. The training data itself consists of eight variables: interval number and target date as indexing variable, 3 power usage measurements in the interval from last three weeks, and 3 outdoor temperature measurements in the interval from last three weeks. More specific details of the parameters that we use to train our ANN prediction mechanism can be found in Table 1.

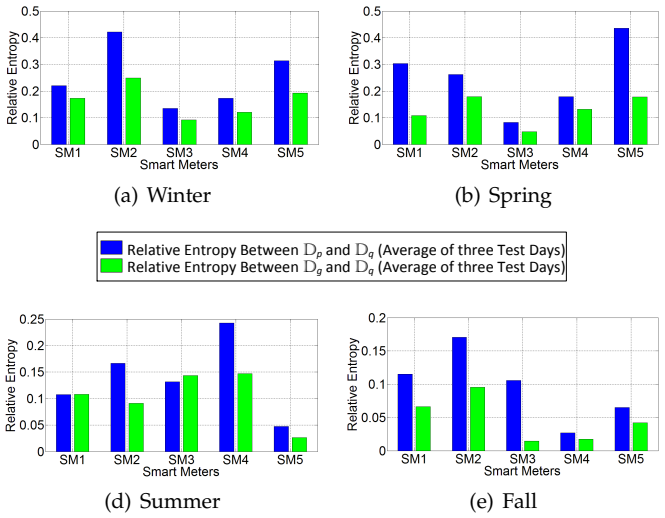
TABLE 1
Neural network training parameters.

Parameter	Value
Number of SMs in cluster (assumed neighborhood)	5
Training period	3 weeks (21 days)
Testing period	3 day
Number of predicted data points a day	288
Number of ANN Inputs	9
ANN Proto	50
Number of ANN hidden layers	3
Number of nodes in each hidden layer	10
Number of ANN output	1
ANN Learn Rule	Ext DBD
ANN transfer mode	Sigmoid
Epoch	288*21=6048
Number of iterations	10^6

5.2 Privacy Implications

To evaluate the privacy implications of Seer Grid, we conduct two different experiments at the SM level. Both the experiments are designed to observe and compare the amount of information that can be inferred from Seer Grid's predicted energy consumption time series data, versus time series data of actual energy consumption. We define $\mathbb{P} = \{p_1, p_2, \ldots, p_n\}$ as the Seer Grid's predicted energy consumption time series and $\mathbb{Q} = \{q_1, q_2, \ldots, q_n\}$ as the actual energy consumption time series, where $p_t, q_t : 1 \leqslant t \leqslant n$ are the energy consumptions for each time interval. We also calculate difference between successive power measurements in $\mathbb{P}$ and $\mathbb{Q}$, symbolized as $\mathbb{D}p = \{dp(1), \ldots, dp(n)\}$ and $\mathbb{D}q = \{dq(1), \ldots, dq(n)\}$, where $dp(i) = p_i - p_{i-1}$ and $dq(j) = q_j - q_{j-1}$. $\mathbb{D}p$ indicates the changes in energy consumption load, which is important to understand privacy leaked through load signatures. We plot $\mathbb{D}p$ and $\mathbb{D}q$ in Figure 6 to visualize if much Seer Grid indeed suppresses load differences. It can be observed that Seer Grid has consistently less changes in energy consumption load throughout the test day, indicating consistent privacy protection. This motivates us to further analyze $\mathbb{D}p$ with respect to $\mathbb{D}q$, and compare privacy leakage of Seer Grid with another well-known protection mechanism in the following two experiments.

Relative Entropy: In our first experiment, we try to quantitatively compare $\mathbb{D}p$ and $\mathbb{D}q$ over four seasons (each



(a) Winter (b) Spring

(d) Summer (e) Fall

Fig. 7. Relative entropy between $\mathbb{D}p$ and $\mathbb{D}q$ compared with relative entropy between $\mathbb{D}g$ and $\mathbb{D}q$, where $\mathbb{D}g$ is the series of differences between successive power measurements in GRN induced energy consumption data.

with 21 training and 3 test days): winter (January 1 to 24), spring (April 1 to 24), summer (July 1 to 24), and fall (October 1 to 24). Let A and B be the probability distributions of $\mathbb{D}p$ and $\mathbb{D}q$, respectively. We use the well-established relative entropy metric (Equation 1) as a non-symmetric measure of difference between the two probability distributions A and B. Due to the premetric property of relative entropy, the larger the relative entropy $D(A\|B)$, the higher the level of protection offered.

In order to understand the level of privacy protection offered by Seer Grid, we compare it's relative entropy with a widely accepted perturbation technique - introduction of Gaussian random noise (GRN) [12]. Introduction of GRN adds or subtracts random values in each reported energy consumption interval, in order to mislead an adversary. However, completely random noise will reduce data utility adversely. Therefore, we use past energy consumption data features to set a level of noise which balances between privacy and utility, as proposed by [12]. Figure 7 shows the relative entropy values for Seer Grid and GRN, for the five smart meters under evaluation. From the results, we observe that the relative entropy of Seer Grid is generally higher compared to relative entropy of GRN, which indicates that Seer Grid may offer better privacy protection.

$$D(A\|B) = \sum_i A(i) \log \frac{A(i)}{B(i)} \tag{1}$$

Clustering: As a second metric to evaluate privacy of Seer Grid, we apply a clustering technique to recover activity information of consumers. Clustering creates groups of similar levels of energy consuming intervals. More number of the clusters can leak more granular information (spikes, switching on/off, and consumption pattern) in each household. In other words, higher number of clusters inferred by an adversary reveals more private information about appliances and activity within the household. We use the Self Organizing Map (SOM) algorithm [39] to create clusters on successive energy difference time series. The interesting aspect of SOM is that it learns to cluster data without supervision. In our application, SOM groups input values into n clusters such that the difference between power consumption values across clusters is minimized. We use the Viscovery tool [40] to apply SOM and calculate the optimum number of clusters for successive energy differences in actual, Seer Grid predicted and GRN perturbed time series (Figure 8). As defined before, $\mathbb{D}q$, $\mathbb{D}p$, and $\mathbb{D}g$ denote the successive differences in actual, Seer Grid and GRN energy time series, respectively. Figure 9 visualizes the clustering done by the SOM algorithm on the three series, for each of the five SMs. Each sub-figure in Figure 9 is clustered into specific distances between the cluster members, where the distance is varied from zero to the maximum in the time series. As evident from Figure 8 and 9, Seer Grid generally has the lowest number of clusters, and thus, reveals least information compared to actual and GRN induced energy time series. Figure 8 also illustrates clusters and distribution of cluster population within each cluster. Because Seer Grid prediction results in a "smoother" pattern, we observe a high population in the low distance clusters.

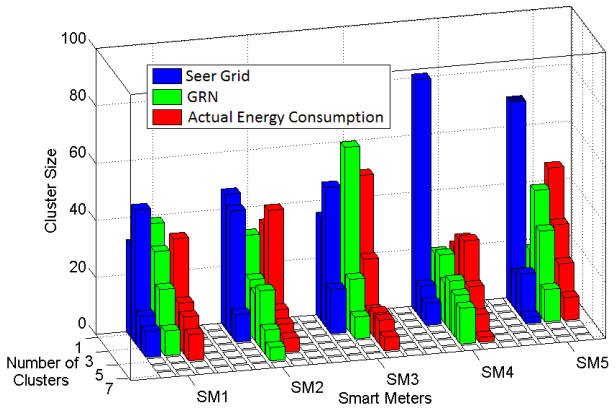


Fig. 8. Number of clusters in $\mathbb{D}q$, $\mathbb{D}p$, and $\mathbb{D}g$, and percentage of distance in each cluster.

Comparison with SARMA [23]: Singh et al. proposed the use of Seasonal Auto Regressive Moving Average (SARMA) for household load prediction. We compare our household level prediction with SARMA by recreating their experiments for our earlier defined test days across four seasons. Figure 10 shows the root mean square error and normalized mean square error percentage in the predicted loads. Higher error percentage means that the prediction is further off from the actual load values, implying more privacy against inference attacks. It is observed that Seer Grid's household level prediction has an overall higher error percentage compared to SARMA, which means Seer Grid offers more privacy than SARMA. However, it should be noted that SARMA was designed to have low error percentage in household level prediction, so as to improve utility. On the other hand, our primary goal for the household level prediction is to improve privacy. However, the error percentage should not be very high, otherwise the cluster level prediction may suffer loss of data utility. We evaluate the utility of the cluster level prediction below, to validate that data utility is in fact not significantly compromised in Seer Grid, even when there exists relatively higher error percentage in household level prediction.

5.3 Data Utility

In this section we empirically evaluate the utility implications of Seer Grid using the well-accepted squared correlation metric [8], [12]. We conduct experiments over four seasons: winter (January 1 to 24), spring (April 1 to 24), summer (July 1 to 24), and fall (October 1 to 24). The results, averaged over the 3 test days, are presented in Table 2. The squared correlation between actual and predicted energy consumption patterns of SM vary between 51.07% and 80.09%, and averages at 62.10% across all 5 SMs. As an example, Figure 11(a) shows the actual and predicted energy consumption pattern for a SM on 22nd January, and Figure 11(b) shows the squared correlation between them. The squared correlation between actual and predicted energy consumption pattern for CH vary between 89.95% and 91.15%, and averages at 90.60%. Figure 11(c) shows the actual and predicted energy consumption pattern for CH on 22nd January, and Figure 11(d) shows the squared correlation between them. Evidently, SM prediction is less

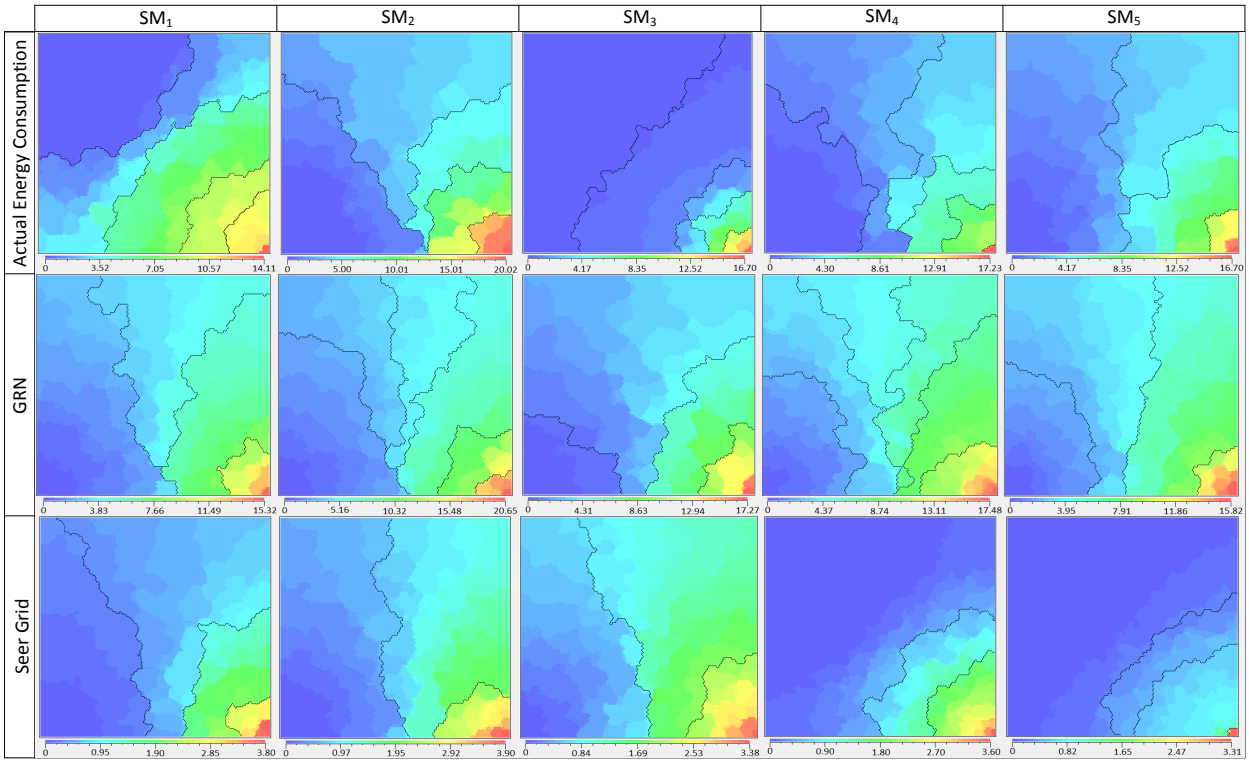


Fig. 9. Cluster forms on $\mathbb{D}q$, $\mathbb{D}p$, and $\mathbb{D}g$. The experiment is performed with 21 training (January 1 to 21) and 3 test days (January 22 to 24). The results are averaged over 3 consecutive test days.

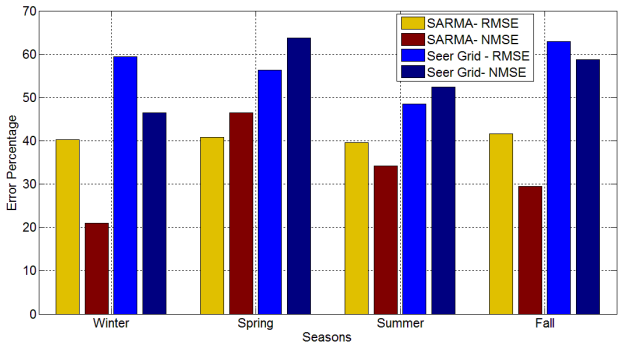


Fig. 10. Root mean square error (RMSE) and normalized mean square error (NMSE) percentage in predicted household loads by SARMA and Seer Grid. Results are averages of the five test SMs.

correlated than CH prediction by a clear margin, as seen in Table 2. We also check the standard deviation of the test days to verify there does not exists any bias. Standard deviation values appear random, without any visible connection with the squared correlation results, leading us to believe that our results are unbiased.

Comparison with Jain and Satish [19]: Jain and Satish proposed the novel use of support vector machines (SVM) to perform short-term load forecasting at cluster level [19]. To better understand how our cluster level prediction would perform, we do a comparative analysis with [19] by recreating their experiments, in the same period of our earlier defined test days across four seasons. Figure 10 shows the maximum percent error and average percent error in cluster level load prediction for Jain and Satish, and Seer Grid. In this case, lower percent error in prediction implies better utility for EC. Seer Grid's cluster level prediction has marginally lower percent error in winter and fall, marginally higher percent error in spring, and a significantly lower percent error in summer. Overall, we can conclude that Seer Grid's utility is similar to [19], if not better. Readers should note that [19]'s SVM based load forecasting at cluster level is only a single level prediction, where the prediction model is trained with actual data from past. Whereas, in case of Seer Grid, the cluster level prediction is primarily based on the household level prediction, which provides better privacy as seen earlier. Therefore, Seer Grid having similar utility as other cluster level prediction schemes is very promising.

6 Discussions

6.1 Smart Meter Performance Analysis

Although Seer Grid uses complex prediction schemes, it does not suffer from significant computational and communicational bottlenecks. As the prediction is once a day event, SMs have an entire day to compute for next day, which should be sufficient even for less powerful computing systems. Reporting the predicted data is also an once-a-day event, and SMs can avoid network congestion if they report using a multiplexed (time, frequency, or code sharing) channel with other SMs. Among all SM privacy preserving frameworks, the most closely resembling (in terms of resource requirements, architecture, and assumptions made) frameworks are based on homomorphic cryptography [41]. So, we compare the computational complexity of Seer Grid

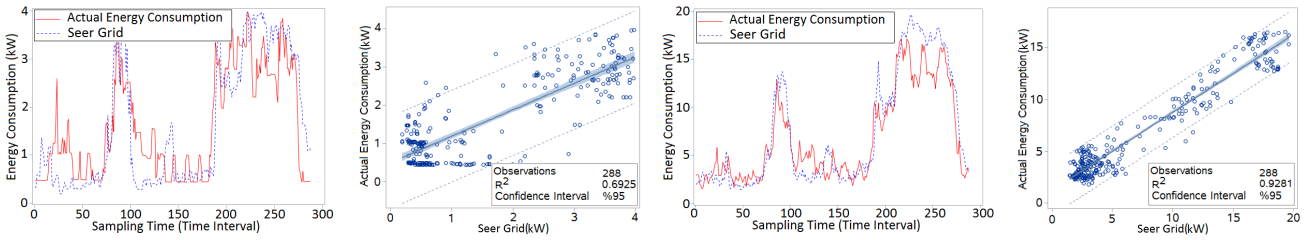


Fig. 11. Exemplary results from 22nd January 2008, showing the correlation between actual and predicted energy consumption patterns at different levels of Seer Grid. (a) Actual and predicted energy consumption patterns for one of the SMs, (b) Correlation between actual and predicted energy consumption patterns for the same SM, (c) Actual and predicted energy consumption patterns for CH, (d) Correlation between actual and predicted energy consumption patterns for CH.

TABLE 2
Squared correlation coefficient (R^2) between predicted and actual energy consumption patterns for each SM and CH, and the standard deviations of the 3 test days.

Season		SM1	SM2	SM3	SM4	SM5	CH
Winter	R^2: Actual vs Predicted	**0.5715**	**0.5529**	**0.7793**	**0.6421**	**0.5772**	***0.9098***
	Three Test Days Standard Deviation	0.1240	0.0618	0.1470	0.0901	0.0187	0.0118
Spring	R^2: Actual vs Predicted	**0.5107**	**0.5627**	**0.8009**	**0.6687**	**0.5799**	***0.9115***
	Three Test Days Standard Deviation	0.0728	0.1236	0.1588	0.0459	0.1868	0.0095
Summer	R^2: Actual vs Predicted	**0.5888**	**0.5341**	**0.6322**	**0.6439**	**0.6528**	***0.8985***
	Three Test Days Standard Deviation	0.1775	0.1366	0.0922	0.0479	0.1855	0.01725
Fall	R^2: Actual vs Predicted	**0.6195**	**0.6025**	**0.6477**	**0.6450**	**0.6072**	***0.9041***
	Three Test Days Standard Deviation	0.0572	0.1412	0.0284	0.0808	0.1074	0.0102

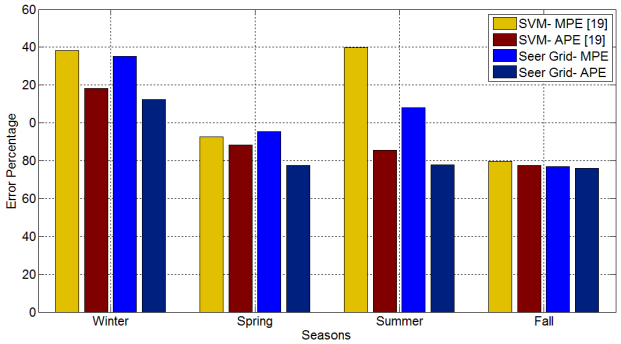


Fig. 12. Maximum percent error (MPE) and average percent error (APE) in cluster level load prediction for Jain and Satish, and Seer Grid. Results are averages of the five test SMs.

versus frameworks based on homomorphic cryptography. Seer Grid's household level prediction through multilayer perceptron has a time complexity of $O(x^2)$ [42], while Paillier cryptography has time complexity of $O(y^3)$ [41]. Therefore, time complexity of homomorphic cryptography based SM privacy frameworks is $O(y^3t)$ while Seer Grid's time complexity is $O(x^2t)$, where x and y are the size of input in Seer Grid and homomorphic (Paillier) cryptosystem [41], respectively, with $x \ll y$, and t is the number of daily samples in both schemes. In other words, Seer Grid's asymptotic time complexity is lower than similar aggregation frameworks based on homomorphic cryptography.

6.2 Implications

The Importance of Two Level Prediction: One may think that only a single level of prediction may achieve the same results as two-levels, but a single level of prediction has some inherent drawbacks. If the prediction is done only at the CH level (where households report their actual consumption to CH), we lose privacy at the SM level. Whereas, if prediction is done only at the SM level, the cluster-wide difference between actual and predicted consumption data will be larger, resulting in data utility loss.

Training Parameters: In our experiments, we took a heuristic approach for determining the training parameters (epochs, iterations, learning rule, etc.) for the ANN used by SMs. The parameters were chosen in such a way that it satisfies our goal of optimizing both privacy and utility of SM data. From the experimental results we observe that the correlation between actual and predicted energy consumption pattern varies moderately across households and seasons. This is primarily because of different characteristics of the training data (actual energy consumption for last 21 days) leading to differently converged prediction model in each SM. In future, we plan to develop a unified prediction framework for the SMs which will analyze characteristics of the training data, and accordingly govern learning rate such that prediction accuracy remains below a privacy preserving threshold with high likelihood. Unlike this work, where all SMs use the same prediction parameters, the unified framework will adapt to the characteristics of local training data of individual SMs. As a result, the convergence in learning will be more uniform across SMs and seasons, thus offer a more stable privacy guarantee.

Privacy due to Uncertainty: Uncertainty in next day's

energy consumption provides user privacy in Seer Grid, which is similar to how uncertainty in location data provides spatio-temporal privacy [43]. The naturally occurring irregularities in consumers' day-to-day schedule results in smoother household prediction patterns (that hides load signatures), which also means that the predicted energy pattern cannot be used to determine temporary house unoccupancies with complete confidence.

Larger Cluster: We consider a very small scale cluster in our experiments, and yet achieve considerably high prediction accuracy at the cluster level. As evident from previous cluster level prediction schemes [44], accuracy tends to dramatically improve with increasing cluster size. Thus, we think our results are highly encouraging for large scale implementation.

6.3 Dishonest and Malfunctioning Smart Meters

SMs are often the target of bad data injection attacks, primarily due to monetary incentives [45] [46]. However, it is critical for ECs to prevent such attacks, not only to avoid financial losses, but also to ensure proper distribution of electricity. Previous efforts in this direction suggested the use of embedded sensors for 'Trusted Metering' [47], having a centralized or dedicated detection system, or a hybrid system of embedded sensors and centralized detection [46]. In Seer Grid, as the CH collects predicted energy consumption data of individual smart meters in advance, existing anomaly detection mechanisms can be effectively applied on the predicted energy consumption data reported by individual smart meters.

In real deployment, SMs and/or communication links can also experience failures, due to which they may be unable to report the predicted energy consumption information. Similar challenge is also faced by existing SMs (and many proposed aggregation schemes), and can be a non-trivial issue to address. If a negligible number of SMs (belonging to the same energy company) are unresponsive, the effects most likely will be unnoticeable. However, if a large number of SMs are unresponsive, the effects can be significant. In Seer Grid, such cases of malfunctioning SMs can perhaps be handled more efficiently than other aggregation schemes, due to the readily available past prediction data. For example, if the next day's predicted energy consumption data is not reported by a SM, the cluster head can simply substitute it with the same week-day's prediction of that SM from last week. The intuition behind this exemplary approach is that households generally have similar usage pattern for each day of the week [48].

6.4 Deployment Barriers

Smart meter deployment presents EC with many logistical, technical and commercial challenges. The primary incentive for ECs to deploy SMs is efficiency and thus savings over time. Conventional SMs, already deployed in many places, perfectly serve this commercial benefit. However, these SMs were not designed to provide privacy for consumers. As a result, any new framework designed to enable consumer privacy will require modification or re-deployment by the EC, which will require additional investment from ECs. Because this new investment does not add any additional efficiency improvements, ECs might be reluctant in deploying any privacy preserving add-ons to existing SMs. This is a major limitation faced by many novel privacy preserving frameworks proposed for smart grids [49]. Cavoukian and Dix [49] pointed out that privacy by design is the best approach. Therefore, deployment of Seer Grid can be easier in new localities (without existing smart metering infrastructure), than to implement in localities where smart metering is already in place. Given that Seer Grid will require additional hardware and software to function, below are the few directions we think can aid deployment:

- *Add-On Service:* ECs can offer SMs with Seer Grid's prediction framework as an add-on service. That is, privacy-aware consumers can opt in for the privacy preserving framework, by paying an one time fee, which would cover the cost of additional hardware and software installation.
- *Off-Loading Computation:* Instead of adding a computing unit (for performing the prediction operations) built inside the SMs, it may be beneficial to off-load the operations to a household computer. For example, the prediction operations can be undertaken by a paired (using low energy communication protocols, such as Bluetooth) smartphone or PC, once per day. The prediction results can be communicated back to the SM for reporting to EC. Also, future upgrades may be easier for consumers, as smartphones and PCs are more frequently upgraded [50].

Alternatively, privacy issues can result in poor SM adoption in privacy-aware communities [51]. By addressing privacy issues in a way that does not hamper utility too much, ECs can increase SM adoption. This can be an incentive for ECs to participate in implementing frameworks like Seer Grid.

Conclusion

We propose Seer Grid, an alternate SGN architecture aimed to reduce the privacy-utility trade-off faced by SMs. As a result of two-level energy load prediction in Seer Grid, there exists high correlation between predicted and actual energy consumption patterns at cluster level, which indicates excellent utility preservation. However, the correlation between predicted and actual energy consumption patterns of individual SM is weak, which indicates good privacy preservation for households. Evaluation results strongly support our proposition of Seer Grid.

Future Work

The goal of this paper is to demonstrate the benefits of using two-level prediction in SGN, using exemplary models. In future, we plan to generalize and formulate the minimization of privacy-utility trade-off in Seer Grid. Also, we plan to quantitatively evaluate Seer Grid against non-intrusive load monitoring attacks, using publicly available data sets, such as thst REDD dataset [52] or the Residential Energy Consumption Survey dataset [53].

ACKNOWLEDGMENTS

This work was supported in part by the Power Systems Engineering Research Center (PSERC) Project S-54.

REFERENCES

[1] Telefnica Digital, "The smart meter revolution - towards a smarter future," Jan 2014.
[2] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, "Occupancy detection from electricity consumption data," in *Proceedings of BuildSys*, 2013.
[3] M. Weiss, A. Helfenstein, F. Mattern, and T. Staake, "Leveraging smart meter data to recognize home appliances," in *IEEE PerCom 2012*.
[4] I. Rouf, H. Mustafa, M. Xu, W. Xu, R. Miller, and M. Gruteser, "Neighborhood watch: security and privacy analysis of automatic meter reading systems," in *ACM CCS 2012*.
[5] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *IEEE Security & Privacy*, vol. 8, no. 1, 2010.
[6] F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *IEEE SmartGridComm 2010*.
[7] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Security and Trust Management*. Springer, 2011.
[8] G. Kalogridis, C. Efthymiou, S. Z. Denic, T. Lewis, R. Cepeda *et al.*, "Privacy for smart meters: Towards undetectable appliance load signatures," in *IEEE SmartGridComm 2010*.
[9] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring," in *ACM CCS 2011*.
[10] X. He, X. Zhang, and C.-C. Kuo, "A distortion-based approach to privacy-preserving metering in smart grids," *Access, IEEE*, vol. 1, 2013.
[11] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor, "Smart meter privacy: A utility-privacy framework," in *IEEE SmartGridComm 2011*.
[12] L. Sankar, S. R. Rajagopalan, S. Mohajer, and H. V. Poor, "Smart meter privacy: A theoretical framework," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, 2013.
[13] R. Dong, A. A. Cárdenas, L. J. Ratliff, H. Ohlsson, and S. S. Sastry, "Quantifying the utility-privacy tradeoff in the smart grid," *CoRR*, vol. abs/1406.2568, 2014.
[14] D. Mashima, "Authenticated down-sampling for privacy-preserving energy usage data sharing," in *IEEE SmartGridComm 2015*.
[15] B. Defend and K. Kursawe, "Implementation of privacy-friendly aggregation for the smart grid," in *Proceedings of ACM SEGS*, 2013.
[16] R. Anderson and S. Fuloria, "On the security economics of electricity metering," in *WEIS*, 2010.
[17] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, 2014.
[18] A. Jain and B. Satish, "Short term load forecasting by clustering technique based on daily average and peak loads," in *IEEE PES*, 2009.
[19] ——, "Clustering based short term load forecasting using support vector machines," in *IEEE Bucharest PowerTech*, 2009.
[20] T. Kucukdeniz, "Long term electricity demand forcesting: An alternative approach with support vector machines," *İÜ Mühendislik Bilimleri Dergisi*, vol. 1, no. 1, 2010.
[21] R. Sevlian and R. Rajagopal, "Short term electricity load forecasting on varying levels of aggregation."
[22] M. Ghofrani, M. Hassanzadeh, M. Etezadi-Amoli, and M. Fadali, "Smart meter based short-term load forecasting for residential customers," in *NAPS 2011*.
[23] R. P. Singh, P. X. Gao, and D. J. Lizotte, "On hourly home peak load prediction," in *IEEE SmartGridComm 2012*.
[24] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *IEEE SmartGridComm 2010*.
[25] M. Jawurek, M. Johns, and K. Rieck, "Smart metering de-pseudonymization," in *Proceedings of ACM ACSAC'27*, 2011.
[26] M. Faisal, A. A. Cardenas, and D. Mashima, "How the quantity and quality of training data impacts re-identification of smart meter users?" in *IEEE SmartGridComm 2015*.
[27] D. Mashima and A. Roy, "Privacy preserving disclosure of authenticated energy usage data," in *IEEE SmartGridComm 2014*.
[28] E. Shi, R. Chow, T. h. Hubert Chan, D. Song, and E. Rieffel, "Privacy-preserving aggregation of time-series data," in *NDSS 2011*.
[29] M. Savi, C. Rottondi, and G. Verticale, "Evaluation of the precision-privacy tradeoff of data perturbation for smart metering," *IEEE Transactions on Smart Grid*, vol. 6, no. 5.
[30] N. C. Truong, J. McInerney, L. Tran-Thanh, E. Costanza, and S. D. Ramchurn, "Forecasting multi-appliance usage for smart home energy management," in *Proceedings of IJCAI*, 2013.
[31] G. E. Box and G. M. Jenkins, "Time series analysis: Forecasting and control," in *Holden-Day series in time series analysis*. Holden-Day, 1976.
[32] J. T. Connor, R. D. Martin, and L. E. Atlas, "Recurrent neural networks and robust time series prediction," *IEEE Transactions on Neural Networks*, vol. 5, no. 2.
[33] J. Daemen and V. Rijmen, *The Design of Rijndael: AES-The Advanced Encryption Standard*. Springer Science & Business Media, 2002.
[34] X. Lu, G. Fan, and R. Hao, "A dynamic token passing mac protocol for mobile ad hoc networks," in *IWCMC 2006*.
[35] S. Thaskani, K. V. Kumar, and G. R. Murthy, "Energy efficient cross-layer design protocol by using token passing mechanism for wsn," in *2011 IEEE Symposium on Computers & Informatics*, pp. 572–575.
[36] M. A. Rahman, M. H. Manshaei, E. S. Al-Shaer, and M. Shehab, "Secure and private data aggregation for energy consumption scheduling in smart grids," *IEEE Transactions on Dependable and Secure Computing*, no. 1.
[37] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Transactions on Information Theory*, vol. 22, no. 6, 1976.
[38] I. Richardson and M. Thomson, "One-minute resolution domestic electricity use data, 2008-2009. colchester, essex: Uk data archive [distributor], october 2010. sn: 6583."
[39] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 13, 1998.
[40] Viscovery SOMine Tool. Data mining suite, eudaptics software gmbh. [Online]. Available: https://www.viscovery.net
[41] A. Bessani and S. Bouchenak, Eds., *Distributed Applications and Interoperable Systems*, ser. Lecture Notes in Computer Science, vol. 9038. Springer, 2015.
[42] M. J. Kearns, *The computational complexity of machine learning*, ser. ACM distinguished dissertations. Cambridge, Mass. MIT Press.
[43] S. Merrill, N. Basalp, J. Biskup, E. Buchmann, C. Clifton, B. Kuijpers, W. Othman, and E. Savas, "Privacy through uncertainty in location-based services," in *Proceedings of IEEE MDM*, 2013.
[44] T. K. Wijaya, J. Humeau, Samuel Franois Roger, M. Vasirani, and K. Aberer, "Individual, Aggregate, and Cluster-based Aggregate Forecasting of Residential Demand," Lausanne, Switzerland, Tech. Rep., 2014.
[45] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," in *Critical Information Infrastructures Security*. Springer, 2009, pp. 176–187.
[46] D. Grochocki, J. H. Huh, R. Berthier, R. Bobba, W. H. Sanders, A. A. Cárdenas, and J. G. Jetcheva, "Ami threats, intrusion detection requirements and deployment recommendations," in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*. IEEE, 2012, pp. 395–400.
[47] M. Raciti and S. Nadjm-Tehrani, "Embedded cyber-physical anomaly detection in smart meters," in *Critical Information Infrastructures Security*. Springer, 2012, pp. 34–45.
[48] J. E. Seem, "Pattern recognition algorithm for determining days of the week with similar energy consumption profiles," *Energy and Buildings*, vol. 37, no. 2, 2005.
[49] A. Cavoukian and A. Dix, *Smart Meters in Europe: Privacy by Design at its Best*. Information and Privacy Commissioner of Ontario, Canada, 2012.
[50] V. S. Venkitachalam, V. Namboodiri, S. Joseph, E. Dee, and C. A. Burdsal, "What, why, and how: Surveying what consumers want in new mobile phones." *IEEE Consumer Electronics Magazine*, vol. 4, no. 2, pp. 54–59, 2015.
[51] D. N.-y. Mah, J. M. van der Vleuten, P. Hills, and J. Tao, "Consumer perceptions of smart grid development: Results of a hong kong survey and policy implications," *Energy Policy*, vol. 49.

[52] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability, San Diego, CA*, vol. 25.
[53] Residential energy consumption survey (recs) - data - u.s. energy information administration (eia). [Online]. Available: https://www.eia.gov/consumption/residential/data/

Arash Boustani received the Bachelor of Computer Engineering and Master of Computer Networking in 2007 and 2009, respectively, from Azad University, Iran. He is currently pursuing the Ph.D. degree in Electrical Engineering and Computer Science at Wichita State University, Wichita, KS, USA. His research is on security and privacy in cyber-physical systems. He is currently working on preserving privacy and increasing availability in cognitive radio based systems in Smart Grid Networks.

Anindya Maiti received the M.S. degree in Electrical Engineering from Wichita State University, USA, in 2014, and the B.Tech. degree in Computer Science and Engineering from Vellore Institute of Technology, India, in 2012 . He is currently pursuing the Ph.D. degree in Electrical Engineering and Computer Science at Wichita State University, USA. His current research interests include cyber-physical system security and privacy.

Sina Yousefian Jazi received his B.S in Industrial Engineering from IAUN, Isfahan, Iran in 2013. He continued his Masters in Industrial Engineering at Wichita State University, Wichita, KS. His main areas of focus are statistical analysis, data analysis, and data visualization.

Murtuza Jadliwala received the Ph.D. degree from the State University of New York at Buffalo, Buffalo, NY, USA, in 2008. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, USA, where he directs the Security, Privacy, Trust, and Ethics in Computing Research Labratory. His current research interests include application, network, and cyber-physical system security and privacy enhancing technologies

Vinod Namboodiri received the Ph.D. degree in electrical and computer engineering from the University of Massachusetts Amherst, Amherst, MA, USA. He is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, USA. His current research interests include designing algorithms and protocols for energy-intelligent and sustainable computing, and designing an effective and scalable communications architecture for smart electric grids. Prof. Namboodiri has served on the Technical Program Committee of IEEE INFOCOM, IEEE SmartGridComm, IEEE GLOBECOM, IEEE ICC, IEEE IPCCC, and IEEE GREENCOM. He is an Active Reviewer for numerous journals and conferences in the mobile computing, green computing, and smart grid areas.