Increasing User Controllability on Device Specific Privacy in the Internet of Things

Waqar Asif\textsuperscript{a},*, Muttukrishnan Rajarajan\textsuperscript{a}, Marios Lestas\textsuperscript{b}

\textsuperscript{a}School of Engineering and Mathematical Sciences, City, University of London, UK.
\textsuperscript{b}Department of Electrical Engineering, Frederick University, Nicosia, Cyprus.

Abstract

With recent advancements in information technology more and more devices are integrated in the Internet of Things. These devices gather significant amount of private information pertinent to a user and while, in some cases it helps in improving the life style of an individual, in others it raises major privacy concerns. This trade-off between utility and privacy is highly dependent upon the devices in consideration and as the utility of the generated data increases, the privacy of an individual decreases. In this paper, we formulate a utility-privacy trade-off that enables a user to make appliance specific decisions as to how much data can be shared. This is achieved by parametrizing the degree of privacy allowed for each device and enabling the user to configure the parameter of each device. We use the smart metering application as the test case scenario for the proposed approach. We evaluate its performance using simulations conducted on the ECO data set. Our results indicate that, the proposed approach is successful in identifying appliances with an accuracy of 81.8% and a precision of 70.1%. In addition, it is demonstrated that device specific changes of the configuration parameters allow the degree of privacy achieved for the particular device and the utility to be well controlled, thus demonstrating the effectiveness of the proposed approach. Moreover, it is shown that, as expected, devices with higher power consumption contribute more to the overall privacy and utility achieved. A comparative study is also conducted and the proposed approach is shown to outperform the existing ElecPrivacy approach by producing a trace that is harder to identify, as reported after testing the Weiss’ and Baranski’s algorithm, both of which are well known Non-Intrusive Load Monitoring algorithms. Finally, it is demonstrated that the addition of noise, which is an integral part of the propose approach, can greatly improve performance.

Keywords: Smart Meter, Privacy, Utility-Privacy trade-off, Mutual Information, Non-Intrusive Load Monitoring.

1. Introduction

Internet of Things (IoT) has been one of the key innovations of the last decade. It eliminates the requirement of having a homogeneous network for ensuring information exchange, thus, enabling multiple devices to communicate with each other. By 2020, it is expected that 24 billion devices will be connected using this innovative technology thus, generating an expected benefit of 1.2 trillion USD for the communication industry [1]. IoT objectives go beyond serving the communications industry, aiming to improve the quality of life. IoT devices are built on a common architecture of repetitive sensing and forwarding of information thus forming time-series data. This time series data keeps the client aware and updated about the changing parameters thus enabling him to react accordingly. A smart health monitoring device enables a doctor to constantly update the record of a patient’s health and later use this to make a detailed analysis [2] Likewise, a smart baby monitor empowers parents with the capability to take pre-emptive measures against the reduced growth of their child with the aid of a constantly updated record of their child’s activity pattern[3][4]. Similarly, smart meters enable consumers to have knowledge on their consumption pattern thus allowing them to adopt a more cost effective lifestyle [5][6]. The benefits of having these IoT devices is only worthwhile if the generated time-series data is kept secure from the access of an adversary.

An adversary can use this data to learn vital information about a person with information spanning from as small as his average heart rate for a certain period of time to as huge as a detailed analysis of a person’s heart rate, location and estimated activity pattern, all using the same health monitor. Such highly private data collection may happen with or without the consent of the user and potentially without informing the users how this data can affect their privacy [7]. The entity collecting the data may also make this data available to third parties such as insurance companies for accessing one’s daily lifestyle to identify if they are maintaining a healthy schedule and then adjusting the premium accordingly [8], marketing agencies for initiating targeted advertisements, law enforcement agencies for detecting illegal activities and burglars for finding out the habits of the occupants of a house [9]. Each IoT
device poses a different set of privacy risks and this has led to different privacy preserving solutions. One common approach that exists in literature is that of completely hiding the generated time-series data, either by using data aggregating methods or randomization techniques [10][11]. These approaches perform well in ensuring privacy but they curtail the benefits of installing these IoT devices by excessively protecting an individual’s data thus, depriving a user from all the possible set of benefits. The problem then is to ensure that the privacy issue does not limit the access to valuable information thus dampening the data economy and at the same time, the privacy door is not widely open for anyone to extract vital information. The need then is to devise a mechanism that empowers the user with complete control over their information thus leaving it at their disposal to decide on the kind of benefits they want in exchange for their private information. Each IoT device harvests a different level of information and while some devices are very clear in what information they gather, others, such as the smart meters, work in disguise.

Smart meters are advertised as devices that sense and forward the consumption pattern of a household but the time-series data that they generate reveals a lot more information. An adversary can extract vital information such as the living pattern of an individual along with thorough information regarding the appliances installed in the household, such as the time of use and the brand of the washing machines, dryer, kettle, stove, freezer and television [9][12]. It is key to highlight here that, smart meters merely store the power consumption for as small as a single second and then forward it to the utility provider on a predefined data forwarding rate. The accuracy and timeliness of this data is the key to all the benefits related to the smart meters and despite the concerns of a user, the utility provider would not be willing to let anyone alter the average power consumption reported between two data forwarding instances. This bounds the user from adding or subtracting any information thus, limiting the choices to merely distorting the available data. Data distortion can be done using various approaches. One approach is to add an external hardware, where the privacy concerned users install an extra battery that is charged and discharged at irregular intervals, thus generating a distorted consumption pattern [13]. Another approach involves amending the smart meter network structure and ensuring neighbourhood-level aggregation of data before relaying it back to the utility provider (Electric company) [12]. These approaches have been reported to perform well in their considered scenarios but the distortion of data at irregular intervals ensures privacy at the cost of the potential benefits of smart meters.

In this paper, we propose a novel data distortion approach that returns the data sharing authority back to the hands of the user. We formulate a utility-privacy trade-off mechanism that enables the user to decide “what percentage”, of “what data”, related to “which device” should be shared with the utility provider. The proposed approach uses the unique signature pattern of each appliance to identify its existence in a harvesting interval and then, based on the choice of the user, hides the signature. We use the term harvesting interval as the time between two data forwarding instances. The proposed approach is tested on the ECO (Electricity Consumptions and Occupancy) data set [14] that provides a unique combination of quality and quantity of electricity consumption. In particular, it contains aggregate electricity consumption data, including real and reactive power for each of the three phases and plug level measurements of selected household appliances. The data is being collected at 1Hz granularity and over a period of 8 months. Despite the size of the data and the huge variation in power consumption pattern of different devices in a household, the proposed approach was successful in correctly identifying appliances with an accuracy of 81.8% and a precision of 70.1%. Simulation results also indicate that the proposed approach is successful in enabling the user to control the privacy of each individual appliance with the aid of a configurable parameter. Furthermore, it is established that the degree of change in the resulting total privacy and utility is proportional to the power consumption of an appliance. The proposed approach is also compared to the existing ElecPrivacy approach [13] where the resultant smart meter readings are tested using two well-known Non-Intrusive Load Monitoring (NILM) algorithms namely: Weiss’ [15] algorithm and Baranski’s algorithm [16]. The proposed approach has been shown to outperform the existing approach by generating traces that are harder to identify thus demonstrating its effectiveness.

The rest of the paper is organized as follows: in Section II we highlight the related work in this field, in Section III, we formulate the considered problem and in Section IV we present the proposed approach. In section V we evaluate its performance using simulations and finally in Section VI we offer our conclusions and future research directions.

2. Related Work

The constantly increasing awareness regarding the relationship between smart meter data and privacy has led to some interesting research in this field of study. The research involves proposed approaches stretching between smart meter data aggregation for introduction of anonymity to fixing the problem at the origination point and altering smart meter data before it is read by the smart meter.

Authors in [12][17][18] propose privacy enhancing approaches using neighbourhood level aggregation and cryptographic protocols. The idea is to use cryptography to secure data from being read by neighbouring smart meters and then aggregating the data of multiple smart meters before it is relayed to the utility provider. Similarly, authors in [19] use data randomization before aggregation, thus completely hiding the energy signature of a single household and then relaying the aggregated data to the
utility provider. The baseline assumption is that the utility provider only needs energy consumption reading for a substation. These approaches introduce privacy to an individual’s data and offer some utility to the utility provider in understanding the average energy consumption pattern of a certain location but it mitigates the benefits that an individual can obtain from his own smart meter reading, as he cannot identify the appliances which contribute most to the total energy consumption thus undermining his ability to achieve energy efficiency.

To address this problem, authors in [20] focus on distorting the data at the origination point. This would minimize the chances of a malicious node’s participation in the smart meter network for both a passive or an active attack and would also enable an individual to keep a constant check of his energy consumption pattern. They propose the use of a stationary Gaussian Markov model for the energy load measurements. They report that privacy-utility trade-off can be optimized through water-filling and for this the privacy mechanism distorts the time-series data off-line after obtaining the whole sequence, thus using memory that increases exponentially with the reduction of the energy harvesting interval. On the other hand, the authors in [21][22] introduce the notion of partial information hiding by introducing uncertainty about individual values in a time-series by perturbing them. A similar notion is presented in [13][23] where it is pointed out that simple data perturbation would be easy to identify by the utility provider as he can identify the actual consumption pattern by installing a similar smart meter at the power origination point. It is thus better to off-load some of the power consumption to batteries at random time intervals. The proposed approach performs well in hiding the consumption pattern at the source but it affects the average consumption per harvesting interval, thus mitigating some benefits of having a smart meter.

Authors in [24][25] highlight the risks of sharing private data and propose appropriate trade-off mechanisms where the user is informed about the usage of his data and an incentive equal to the privacy cost is offered in return. In [26] a new Multi-party Access Control (MPAC) model was proposed that uses Nash Equilibrium to obtain a unique solution for sharing data with multiple parties and identify an appropriate incentive. The difficulty with these approaches is that, they are highly influenced by the truthfulness and greediness of the user which can entice them into affecting the originality of the data. This was dealt with in [27] where, the authors addressed the issue of participation of consumers in data analytical projects using a game theoretic model. The main idea is to encourage contribution into these studies at an individual self chosen level of precision, with the condition that the provided data is authentic. The analyst sets a requirement for data precision, which has to be met by the provided data.

In this paper, we formulate a utility-privacy trade-off problem around which we propose a solution that enables users to decide what amount of information regarding which appliance is shared. This enables the user to bargain with the utility provider about the information he is willing to share and the benefits he can get in return while maintaining the same average consumption per harvesting interval. The next section explains in detail the problem formulation and later we explain the proposed approach.

3. Problem Formulation

Time series data extracted using a smart meter is critical for ensuring efficient load management and is perilous if in the hand of an adversary. An adversary can use this data to extract key private information about a user. In this section, we formulate the utility-privacy trade-off using a mathematical framework, where the utility of the data refers to the amount of meaningful information that can be extracted from it.

We assume that the harvesting interval of a smart meter is set by the utility provider and remains constant. During each harvesting interval, $t$ time-series values are stored by the smart meter. These time-series values are collectively represented as a set $X^T$, where $T$ represents the data forwarding instance. The time-series values in $X^T$ can be partitioned into two subsets, the public set $S$ which the user is willing to share and the private data set $Q_r$ which the user wants to hide, where $X^T = (S \cup Q_r)$ and $(S \cap Q_r) = \phi$. The private data set $Q_r$ comprises of $r$ subsets $(Q_1 \cup Q_2 \cup \ldots \cup Q_n) = Q_r$, where each subset refers to a unique appliance in a household. Each unique appliance works for a different length of time, therefore, each data set $X^T$ contains unique length subsets $|Q_1|, |Q_2|, \ldots, |Q_r|$ where, $|.|$ represents the cardinality of a set. With such a data set, the requirement for maintaining the actual average consumption per harvesting interval is fulfilled by merely spreading the private data set $Q_r$ across the public set $S$ thus producing a new time-series data $X^T_r$, where $(X^T \cap Q_r) = \phi$. This data can then be released to the provider for ensuring privacy.

We build on the information leakage approach proposed in [28] to identify the mutual information between the private and the randomized data, $I(Q_r; \tilde{X}^T) = H(Q_r) - H(Q_r|\tilde{X}^T)$, where $H(.)$ and $H(.|)$ denote the entropy and the conditional entropy respectively. Here, entropy is referred to as the expected amount of information contained.

An optimal utility-privacy trade-off approach should minimize the expected distortion between the original and the randomized data while ensuring a privacy constraint $\epsilon_r$ set by the user for each individual secure data set $Q_r$, thus forming the following optimization problem:

$$O = \min_{\mu(X^T|X^T, Q_r)} \mathbb{E}[d(X^T, \tilde{X}^T)]$$

subject to $I(Q_r; \tilde{X}^T) \leq (1 - \epsilon_r) \times I(Q_r; X^T), \forall r$
Mutual information has been widely used as a privacy metric due to two main reasons: its well understood properties and its strong correlation with the well-known $\epsilon$-differential privacy [29][30][31]. Differential privacy guarantees that two adjacent databases which differ in only one entry are statistically indistinguishable [32]. This guarantee is particularly effective in making individuals comfortable in contributing personal information in to a database. Consider a database $D^n = (D_1, \ldots, D_n)$ that returns a query response $Y$ according to a random mechanism $P_{Y|D^n}$ and let $D^{-i}$ denote the set of database entries excluding $D_i$ then:

**Definition 1.** (($\epsilon,\delta$)-Differential Privacy (DP) [33]). A randomized mechanism $P_{Y|D^n}$ satisfies $(\epsilon,\delta)$-differential privacy if for all neighbouring database instances $d^n$ and $\tilde{d}^n$

$$P_{Y|D^n = d^n}(\epsilon,\delta) \approx P_{Y|D^n = \tilde{d}^n}(\epsilon,\delta)$$

Where a randomized mechanism $P_{Y|D^n}$ satisfies $\epsilon$-differential privacy if it satisfies $(\epsilon,0)$-differential privacy. Here, a smaller value of $\epsilon$ represents a higher level of privacy. An adversary if generates a query $Y$ for a sufficiently small $\epsilon$ in $\epsilon$-differential privacy, then the query response would be insufficient to distinguish between the two databases $D^n$ and $D^{-i}$. In order to build an information theoretical understanding of the phenomenon of $\epsilon$-differential privacy it is combined with mutual information thus yielding Mutual Information Differential Privacy [34][35].

**Definition 2.** (Mutual-Information Differential Privacy (MI-DP)[35]). A randomized mechanism $P_{Y|D^n}$ satisfies $\epsilon$-mutual information differential privacy if:

$$\sup_{(i,P_{\tilde{D}})} I(D_i;Y|D^{-i}) \leq \epsilon \text{ bits}$$

The interpretation of $\epsilon$-differential privacy into information theoretic science relaxes the strict privacy constraints of $\epsilon$-differential privacy thus sandwiching Mutual Information Differential Privacy between $\epsilon$-differential privacy and $(\epsilon,\delta)$-differential privacy [35]:

$$\epsilon - \text{DP} \geq \text{MI} - \text{DP} \geq (\epsilon,\delta) - \text{DP}$$

The claim herein is that MI-DP is a weaker constraint than $\epsilon$-DP but a stronger constraint than $(\epsilon,\delta)$-DP. That is if a mechanism holds $\epsilon$-MI-DP then it holds $(\epsilon,\delta)$-DP [35]. Where, a lower value of $\epsilon$ represents a higher level of privacy [34]. The rationale build using the aforementioned privacy definitions states that the lower the mutual information between the private dataset $Q_r$ and the time-series dataset $\tilde{X}^T$ in the optimization problem $O$, the lower the mutual information between the removed private dataset entries $D_i$ and the query response $Y|D^{-i}$ in Eq 3. Therefore, a higher $\epsilon_r$ would yield a lower $\epsilon$ for $\epsilon$-MI-DP thus yielding a higher privacy level.

**4. Proposed Approach**

In this section, we describe the rationale behind the proposed approach with relevance to the optimization problem $O$. The proposed approach is built on the basis that, smart meters read consumption data for each household after every small time interval and this can be as small as a single second. The smaller the time interval the finer the granularity of the extracted data and thus the bigger the privacy risk. These meters store these readings for a certain time and then relay it to the central unit or neighbouring nodes for data aggregation and forwarding. This is merely a design choice and we assume for simplicity that, the user is sending data directly to the utility provider. It is worth mentioning that, this assumption only exists for ease of understanding and the proposed approach would work equally well when data is relayed to neighbouring nodes for data aggregation.

The proposed approach comprises of two steps: the data acquisition and pre-processing step and the data distortion step. The data acquisition step builds a signature database for all the appliances using the peak and rms voltages and currents along with the real and reactive power. The data distortion step uses the output of the first step to identify the appliances and then alter the smart meter signature. The granularity of the extracted data and the continuous store and forward mechanism of a smart meter requires the smart meter reading to be changed on a real time basis. This motivates the need of an Energy Management Unit (EMU) that keeps a constant check on the energy requirement of the household and controls an energy storage unit, a battery in our case, that can either store or provide energy on the desired time instance as shown in Fig 1. The EMU maps the energy requirement of each time instance onto the available appliance signatures in search of a positive identification. In the event of a positive identification at a particular instance, the corresponding value is then hidden based on the user defined $\epsilon_r$ value. In case an appliance identified at time $t$ needs to be hidden then, the EMU calculates the average consumption $\gamma$ of the household up till time $t$ and uses this value to decide if the battery needs to be charged or discharged. The objective is to ensure that the smart meter is only providing the average consumption $\gamma$. The rest is offered by the battery. This differentiates the average consumption reported by a smart meter from the actual household consumption per harvesting interval. The difference between the two averages is realized by altering the average consumption $\gamma$ of future instances using a noise addition approach explained in detail in the next section.

**4.1. Step 1: Data Acquisition and Pre-processing**

Electrical appliances differ from each other based on their electrical signatures. An appliance can be of resistive, capacitive, inductive or a predominance type based on its internal circuitry. For instance, a kettle is almost...
purely resistive whereas, a fan can be predominantly inductive. A resistive appliance is simple, as it uses the available current and voltage to produce power. The capacitive and inductive circuitries on the other hand act differently. An inductive circuit advances the current with respect to the voltage and the opposite happens in case of a capacitive circuit. As a consequence the power transferred to an appliance is less than the simple product of voltage and current that is applied across them. To better understand this concept, the terms active and reactive power are used. In general, appliances work through the active power, while the reactive power is due to the presence of an inductive or capacitive component of an appliance. The combination of the active and reactive power forms another key parameter known as the power factor. The combination of the active power, reactive power and the power factor define the signature parameters of an appliance and these are explained in detail later in this section. For steady state linear time-invariant loads, complex power can be calculated from voltage, current and phase angles as follows [36]:

$$ S = \frac{1}{2} V_m I_m e^{j(\theta_V - \theta_I)} = P + jQ $$

(5)

$$ Pf = \frac{P}{|S|} $$

(6)

where variables $V_m$ and $I_m$ are maximum values of voltage and current respectively and variables $\theta_V$ and $\theta_I$ are phase angles of voltage and current respectively. Here, $P$ represents the real power, $Q$ the reactive power and $Pf$ represents the power factor of the system.

The computation of power is different for non-linear loads and is represented using Fourier series expansion. The number of terms in the expansion represents the dimensions of the features and each coefficient corresponds to the current and voltage in each of the harmonics. Real and reactive power can thus be computed by [36]:

$$ I_n = Z_n^{-1} V_n = Y_n V_n $$

(7)

where the real power is defined as:

$$ P = \sum_{n=0}^{N} P_n = V_0 I_0 + \sum_{n=1}^{N} V_n I_n \cos(\theta_{V_n} - \theta_{I_n}) $$

(8)

and the reactive power is defined as:

$$ Q = \sum_{n=1}^{N} V_n I_n \sin(\theta_{V_n} - \theta_{I_n}) $$

(9)

where $n$ is the frequency number, $Y_n$ and $Z_n$ are the admittance and impedance matrices respectively. $V_0$ and $I_0$ are the DC voltage and current respectively and $V_n$ and $I_n$ are their effective $n^{th}$ harmonic.

The power, current and voltage values are then used to differentiate between different appliances and each of the unique consumption pattern is referred to as the signature of an appliance. An appliance signature can be obtained using various mechanism. One common approach requires the installation of smart plugs. Each appliance in a household is powered using a separate smart plug which then connects the appliance with the EMU. An EMU can then keep a constant track of the power utilization pattern of a household. A secondary and more feasible approach is the use of an outlier/anomaly detection algorithm at the EMU. Anomaly detection algorithms such as Z-score analysis [37], the modified Z-Score [38] or Kurtosis computation [39] enable the EMU with the capability of identifying the installation of a new appliance in the household which was not registered with the EMU earlier. With the help of these algorithms, the EMU keeps a look-out for any anomality in the power utilization pattern. An anomaly once identified is reported by the EMU and can be labelled by the user as an appliance that was powered ON at that particular instance. In this work, we use the latter approach for storing appliance signatures. The underline assumption is that the EMU would detect the anomaly in the power consumption pattern and the user is aware of the appliance that was powered ON at that particular instance. Once all the appliances in a household are registered, the user would not have to give any further inputs to the EMU. The appliance signatures obtained using the aforementioned approach has the following features:

- **Real power**: This is the first component that differentiates between appliances with dissimilar consumptions.

- **Power factor**: Appliances with similar consumption are differentiated based on the power factor, which can differ between appliances of resistive, capacitive and inductive types.

- **Peak current**: Appliance signatures also store the peak current which is the maximum amount of current the appliance requires before reacting. This relates directly to the appliance circuit and can help identify appliances of similar type but different manufacturers.

- **Peak voltage**: This identifies the maximum voltage required by the appliance to start. This along with the peak current forms the initial spike that is used to identify the start of a new appliance.

- **RMS current**: After the initial peak current, the current requirement of an appliance follow a time-varying sinusoidal function. The effective value of
this current function is referred to as the RMS current.

- **RMS voltage**: Similar to the RMS current, after the initial peak voltage, the voltage requirement of an appliance follows a time-varying sinusoidal function and the effective value of this voltage function is referred to as the RMS voltage.

The signature data base generated using the aforementioned features is later used in the next step to identify the appliances in the smart meter acquired time-series data and once the appliances are identified, their signatures are scattered using a novel battery charging and discharging pattern.

### 4.2. Step 2: Data Distortion

In this section we explain in detail how the proposed approach introduces distortion in the smart meter data on run-time basis.

We consider the smart meter harvested data as a finite value data set $X^T$ where the cardinality $|X^T|$ depends on the harvesting interval $\beta$ and the granularity with which the smart meter stores the consumption reading. The EMU maps each value $X^T_i \in X^T$ with the appliance signature database $Q_r$. A positive match between the signature database $Q_r$ and $X^T_i$ initiates the data distortion process. The EMU compares $\epsilon_r$, the user defined privacy level, with $\mu$ the constantly updated threshold. Here, $\{\epsilon_r | \epsilon_r \in \mathbb{R}, 0 \leq \epsilon_r \leq 1\}$ with $\epsilon_r = 1$ representing maximum privacy and $\epsilon_r = 0$ representing minimum privacy. The constantly updating threshold, $\mu = \frac{Eh_r}{(Et_r + Ec_r)}$ is the ratio between the total hidden signature $Eh_r$, the total energy consumed by an appliance $Et_r$ and the energy required by the appliance $Ec_r$. The comparison between $\epsilon_r$ and $\mu$ can yield two possible outcomes: 1) $\epsilon_r < \mu$, 2) $\epsilon_r \geq \mu$. The first outcome means no distortion is required and thus the signature should be left un-altered. The second outcome on the other hand, requires the signature to be altered and this is achieved by repeatedly charging or discharging the battery in a manner that the smart meter reads a value close to the average consumption value of a household. The average consumption is represented by $\gamma$ and it is calculated using:

$$\gamma = \sum_{i=0}^{T} X^T_i + N_i$$  \hspace{1cm} (10)

Here, $N$ represents noise and it is explained in detail later in this section. The average consumption $\gamma$ is later compared with the current energy requirement of the household to decide between the charging or discharging cycles of the battery. In an ideal scenario, if $X^T_i < \gamma$ the EMU would discharge the battery, requesting an amount of $X^T_i - \gamma$ Joules from the battery and fulfilling the total energy requirement by requesting $\gamma$ joules from the utility provider through the smart meter. On the flip side, if $X^T_i > \gamma$ the EMU would charge the battery, requesting an amount of $\gamma$ from the utility provider and charging the battery using the surplus amount of $\gamma - X^T_i$ joules. The constant charging and discharging of the battery will result in the constant difference between the average energy consumption reported by the smart meter and the actual average consumption per harvesting interval. This difference from the average consumption per harvesting interval is stored in $R_m$ and is later used to form noise $N$ for Eq 10.

#### 4.2.1. Battery Discharge ($X^T_i > \gamma$)

Every battery is upper and lower bounded by its maximum and minimum storage capacity respectively and this limits it from providing $X^T_i - \gamma$ joules for all the desired incidences. To overcome this limitation, the EMU makes a decision $d_d = \max((X^T_i - \gamma), (B_c - B_{min}))$ that will result in two possible cases. Here $B_c$ is the current energy of the battery, $B'_c$ is the remaining energy in the battery and $B_{min}$ is the minimum energy required to keep a battery alive.

- Case 1: $(X^T_i - \gamma) > (B_c - B_{min})$

  When the difference in the energy requirement of a household and the running average is greater than the capacity of the battery then, the battery is completely drained and the remaining energy requirement is fulfilled by the electricity provider which dictates the smart meter reading. The energy required from the battery and the smart meter are calculated using the following equations:

  $$\hat{X}^T_i = X^T_i - (B_c - B_{min})$$  \hspace{1cm} (11)

  $$R_m = R_m + (B_c - B_{min})$$  \hspace{1cm} (12)

  $$B'_c = B_c - B_{min}$$  \hspace{1cm} (13)

- Case 2: $(X^T_i - \gamma) < (B_c - B_{min})$

  When the difference in energy requirement of a household and the running average is less than the capacity of the battery then, only the average consumption $\gamma$ is fulfilled by the electricity provider which dictates the smart meter reading and the rest is fulfilled from the battery. The energy requirements are calculated using the following equations:

  $$\hat{X}^T_i = X^T_i - (X^T_i - \gamma)$$  \hspace{1cm} (14)

  $$R_m = R_m + (X^T_i - \gamma)$$  \hspace{1cm} (15)

  $$B'_c = B_c - (X^T_i - \gamma)$$  \hspace{1cm} (16)

#### 4.2.2. Battery Charging ($X^T_i < \gamma$)

Every time the requested energy is less than the average $\gamma$, the EMU will use extra energy from the smart meter and charge the battery. This charging process will be limited by the maximum energy $B_{max}$ of the battery and to adapt accordingly to the current capacity of the battery and once the appliances are identified, their signatures are scattered using a novel battery charging and discharging pattern.

$$\epsilon_r \geq r$$  \hspace{1cm} (10)

Here, $N$ represents noise and it is explained in detail later in this section. The average consumption $\gamma$ is later compared with the current energy requirement of the household to decide between the charging or discharging cycles of the battery. In an ideal scenario, if $X^T_i > \gamma$ the EMU would charge the battery, requesting an amount of $\gamma$ from the utility provider and charging the battery using the surplus amount of $\gamma - X^T_i$ joules. The constant charging and discharging of the battery will result in the constant difference between the average energy consumption reported by the smart meter and the actual average consumption per harvesting interval. This difference from the average consumption per harvesting interval is stored in $R_m$ and is later used to form noise $N$ for Eq 10.

#### 4.2.1. Battery Discharge ($X^T_i > \gamma$)

Every battery is upper and lower bounded by its maximum and minimum storage capacity respectively and this limits it from providing $X^T_i - \gamma$ joules for all the desired incidences. To overcome this limitation, the EMU makes a decision $d_d = \max((X^T_i - \gamma), (B_c - B_{min}))$ that will result in two possible cases. Here $B_c$ is the current energy of the battery, $B'_c$ is the remaining energy in the battery and $B_{min}$ is the minimum energy required to keep a battery alive.

- Case 1: $(X^T_i - \gamma) > (B_c - B_{min})$

  When the difference in the energy requirement of a household and the running average is greater than the capacity of the battery then, the battery is completely drained and the remaining energy requirement is fulfilled by the electricity provider which dictates the smart meter reading. The energy required from the battery and the smart meter are calculated using the following equations:

  $$\hat{X}^T_i = X^T_i - (B_c - B_{min})$$  \hspace{1cm} (11)

  $$R_m = R_m + (B_c - B_{min})$$  \hspace{1cm} (12)

  $$B'_c = B_c - B_{min}$$  \hspace{1cm} (13)

- Case 2: $(X^T_i - \gamma) < (B_c - B_{min})$

  When the difference in energy requirement of a household and the running average is less than the capacity of the battery then, only the average consumption $\gamma$ is fulfilled by the electricity provider which dictates the smart meter reading and the rest is fulfilled from the battery. The energy requirements are calculated using the following equations:

  $$\hat{X}^T_i = X^T_i - (X^T_i - \gamma)$$  \hspace{1cm} (14)

  $$R_m = R_m + (X^T_i - \gamma)$$  \hspace{1cm} (15)

  $$B'_c = B_c - (X^T_i - \gamma)$$  \hspace{1cm} (16)
battery, the EMU will make the decision $d_{c} = \max((\gamma - X_{i}^T), (B_{\text{max}} - B_{c}))$. This decision would result in the following two cases:

- **Case 1:** $(\gamma - X_{i}^T) > (B_{\text{max}} - B_{c})$
  
  When the difference between the average energy and the energy required by the household is greater than the capacity of the battery, then the decision of the amount of energy to be forwarded to the battery and the energy fulfilled through a smart meter are made as follows:

  $\hat{X}_{i}^T = X_{i}^T + (B_{\text{max}} - B_{c})$  \hspace{1cm} (17)
  
  $R_{m} = R_{m} - (B_{\text{max}} - B_{c})$  \hspace{1cm} (18)
  
  $B'_{c} = B_{c} + (\gamma - X_{i}^T)$  \hspace{1cm} (19)

- **Case 2:** $(\gamma - X_{i}^T) < (B_{\text{max}} - B_{c})$
  
  When the difference between the average energy and the requested energy is less than the maximum available capacity of the battery, then the EMU will extract the average energy from the smart meter and forward the excess energy to the battery as follows:

  $\hat{X}_{i}^T = X_{i}^T + (\gamma - X_{i}^T)$  \hspace{1cm} (20)
  
  $R_{m} = R_{m} - (\gamma - X_{i}^T)$  \hspace{1cm} (21)
  
  $B'_{c} = B_{c} + (\gamma - X_{i}^T)$  \hspace{1cm} (22)

Notice that the battery charge and discharge mechanisms mentioned in this section update the battery status along with the required energy at every time instance $i$ where, $i$ can be as small as a single second but this does not mean that the battery is expected to switch between charging and discharging state at every time instance. Any appliance running in a household requires an initial power spike to start its operation, thus generating the peak voltage and peak current pattern, which is then followed by a constant RMS current and voltage requirement that keeps the appliance running. This complete power consumption pattern consumes multiple time instances, usually the order to hundreds thus eliminating the need for the battery to change states at every time instance. The charging and discharging pattern of a battery is highly dependent upon the type of battery being used. These different patterns compliment the noise addition objective of the proposed approach therefore, in this paper we do not specify a particular battery type.

4.2.3. **Noise Addition ($N$)**

The calculation of noise $N$ is inspired by the Haar Wavelet transform (HWT) [40] and is used to accommodate the deviation between the actual average consumption of the household during a harvesting interval and the average consumption being reported by the smart meter after distortion. The difference between the averages is spread across the future time indices of a harvesting interval to influence the calculation of $\gamma$. The addition of noise ensures that the smart meter reports actual average consumption of a household.

The Noise $N$ is calculated by using multiple random pairs $[L_{m}, -L_{m}]$ that form the leaf nodes of a tree, where $0 < L_{m} < B_{\text{max}}$ and the sum of all the leaf nodes is zero. These leaf nodes are used to generate the wavelet coefficients $c$ for each of the internal nodes, such that $c = (a_{1} - a_{2})/2$, where $a_{1}$ and $a_{2}$ are the average values for the leaves in the left and right subtree. The depth of the tree is a design choice with a smaller depth yielding more benefits. After processing all the internal nodes, an additional coefficient, referred to as the base coefficient $c_{0}$ is generated by taking the mean of all leaves nodes. The base coefficient will always result in a zero due to the selection of the random pairs $[L_{m}, -L_{m}]$.

\[
\begin{align*}
\text{Noise} N = \text{average consumption being reported by the smart meter} \\
\text{Noise} N = \text{average consumption of a household.}
\end{align*}
\]

The Noise $N$ is calculated by using multiple random pairs $[L_{m}, -L_{m}]$ that form the leaf nodes of a tree, where $0 < L_{m} < B_{\text{max}}$ and the sum of all the leaf nodes is zero. These leaf nodes are used to generate the wavelet coefficients $c$ for each of the internal nodes, such that $c = (a_{1} - a_{2})/2$, where $a_{1}$ and $a_{2}$ are the average values for the leaves in the left and right subtree. The depth of the tree is a design choice with a smaller depth yielding more benefits. After processing all the internal nodes, an additional coefficient, referred to as the base coefficient $c_{0}$ is generated by taking the mean of all leaves nodes. The base coefficient will always result in a zero due to the selection of the random pairs $[L_{m}, -L_{m}]$.

\[
\begin{align*}
\text{Example: Fig 2 illustrates a HWT with 8 entries $v_{1}, \ldots, v_{8}$, Each number in a circle shows the value of a wavelet coefficient. The base coefficient $c_{0}$ equals the mean of the entries. The coefficient $c_{1}$ is the mean difference between the averages of the left and right subtree $c_{1} = (45 - (-45))/2$. Once the leaf values are spread, the base coefficient is updated with $R_{m}$ as: $c_{0} = c_{0} + L_{m}/2$, where $l$ is the depth of the tree. In order to reduce the chances of an adversary identifying the amount of addition into a leaf node, Laplacian noise is added into each branch coefficient. This Laplacian noise is drawn from the Laplacian distribution with the probability density function $p(x|\lambda) = \frac{1}{2\lambda}e^{-|x|/\lambda}$, where $\lambda$ is a design choice. The addition of Laplacian noise will affect the average consumption of a household reported by the smart meter. This is taken care by the use of a mean subtraction procedure. The mean subtraction procedure divides all but the base coefficient into disjoint sibling groups such that each maximal set of coefficients have the same parent in the tree. For example, in Fig 2, the tree can be divided into three sibling groups: $\{c_{2}, c_{3}\}$, $\{c_{4}, c_{5}\}$, and $\{c_{6}, c_{7}\}$. Next for each sibling group, the noise mean is computed and subtracted from each coefficient in the group, thus ensuring no change in average harvesting interval energy.
\end{align*}
\]

Figure 2: Haar Wavelet Transform.
The intuition is that, after the mean subtraction, the noise coefficient in the same siblings group will sum up to zero. Finally, $N_i$ for all future instances is calculated as following:

$$N_i = c_0 + \sum_{j=1}^{l} (g_j \cdot c)$$

where $g_j$ equals 1 ($-1$) if $v$ is in the left(right) subtree of $c$.

5. Performance Evaluation

In this section, we evaluate the performance of the proposed user controllable device specific privacy preserving approach by testing it on the ECO data set and conducting simulations on Matlab. The ECO data set presents real sensor data for 6 households in Switzerland over a period of 8 months (June 2012 to January 2013). It contains measurements of real and reactive power for each of the three phases in a household and also provides the plug-level data at 1 Hz frequency. We test the proposed approach on household 1 from the ECO data set which has 6 different appliances namely: PC, dryer, washing machine, kettle, fridge and freezer.

We evaluate the accuracy of appliance identification by comparing the results generated by the proposed approach with the available plug level data.

Table 1 presents the confusion matrix of the proposed approach for identifying the number of times in which an appliance was either confused with other appliances or was detected correctly. The number of times an appliance was correctly identified while being powered ON is referred to as True Positive (TP), whereas, if an appliance was not working and was incorrectly reported as being powered ON, is referred to as False Positive (FP). Similarly, the number of times an appliance was classified OFF when it was actually ON is referred to as False Negative (FN) and the fraction of times an appliance was correctly classified as being ON while it was actually ON is referred to as the true Positive Rate (TPR) and the fraction of times the appliance was incorrectly classified OFF is referred to as False Negative Rate (FNR) and the True Negative Rate (TNR) are defined as shown below:

$$TPR = \frac{TP}{(TP + FN)} \quad FPR = \frac{FP}{(FP + TN)}$$

$$FNR = \frac{FN}{(TP + FN)} \quad TNR = \frac{TN}{(FP + TN)}$$

The table also presents the Precision $= \frac{TP}{TP + FP}$ and Accuracy $= \frac{TP + TN}{TP + TN + FN + FP}$. The accuracy and precision of the proposed approach is highly affected by the TPR value which was found to be 69%. This is due to the fact that the proposed approach uses the smart meter reading to identify individual appliances and compares the results with the available plug level information for each individual appliance. The plug level information of a device contains operation cycles for an individual appliance only whereas, the smart meter data contains energy utilization of multiple devices running concurrently and this makes the identification of individual appliances difficult. A higher power consuming device overshadows devices that consume less power thus reducing the accuracy of the proposed approach. In the considered household, the Washing Machine, PC and Freezer all work on the same phase, where a Washing Machine consumes the most amount of power of around 2300±100 Watts and the Freezer consumes a peak power of 80 watts followed by an
Table 2: Percentage participation of each appliance with varying $\epsilon_r$ for individual devices.

<table>
<thead>
<tr>
<th>Appliance</th>
<th>$\epsilon_r=0$</th>
<th>$\epsilon_r=0.25$</th>
<th>$\epsilon_r=0.5$</th>
<th>$\epsilon_r=0.75$</th>
<th>$\epsilon_r=1$</th>
<th>$\epsilon_r=2$</th>
<th>$\epsilon_r=2.25$</th>
<th>$\epsilon_r=0.5$</th>
<th>$\epsilon_r=2.75$</th>
<th>$\epsilon_r=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fridge (e1)</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
</tr>
<tr>
<td>Dryer (e2)</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
</tr>
<tr>
<td>Kettle (e3)</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
</tr>
<tr>
<td>Washing Machine (e4)</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
</tr>
<tr>
<td>PC (e5)</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Freezer (e6)</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
</tr>
<tr>
<td>Others</td>
<td>22.15%</td>
<td>24.67%</td>
<td>27.18%</td>
<td>29.70%</td>
<td>32.19%</td>
<td>22.15%</td>
<td>22.36%</td>
<td>22.57%</td>
<td>22.79%</td>
<td>23.00%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Appliance</th>
<th>$\epsilon_r=0$</th>
<th>$\epsilon_r=0.25$</th>
<th>$\epsilon_r=0.5$</th>
<th>$\epsilon_r=0.75$</th>
<th>$\epsilon_r=1$</th>
<th>$\epsilon_r=2$</th>
<th>$\epsilon_r=2.25$</th>
<th>$\epsilon_r=0.5$</th>
<th>$\epsilon_r=2.75$</th>
<th>$\epsilon_r=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fridge (e1)</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
<td>8.61%</td>
</tr>
<tr>
<td>Dryer (e2)</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
<td>0.85%</td>
</tr>
<tr>
<td>Kettle (e3)</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
<td>4.89%</td>
</tr>
<tr>
<td>Washing Machine (e4)</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
<td>29.70%</td>
</tr>
<tr>
<td>PC (e5)</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.01%</td>
</tr>
<tr>
<td>Freezer (e6)</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
<td>33.79%</td>
</tr>
<tr>
<td>Others</td>
<td>22.15%</td>
<td>24.67%</td>
<td>27.18%</td>
<td>29.70%</td>
<td>32.19%</td>
<td>22.15%</td>
<td>22.36%</td>
<td>22.57%</td>
<td>22.79%</td>
<td>23.00%</td>
</tr>
</tbody>
</table>

RMS power of 40 ± 5 Watts, therefore, it is highly likely that when a Washing Machine is in operation, it is difficult to identify the operation status of a Freezer. This goes without saying that the Freezer is in operation throughout the day and a single operation cycle of a Washing Machine overshadows multiple operation cycles of the Freezer thus resulting in a lower TPR value of 69%. In such a scenario, the proposed approach performs well in identifying the appliances correctly with an accuracy of 80.1% and a precision of 70.1%.

In the next set of experiments, we investigate the change in percentage of the identifiable appliance signature as $\epsilon_r$ is increased for each appliance individually. Each appliance has a distinctive power consumption pattern thus generating a unique signature as reported in Fig 3. Adequate changes to this signature pattern would render the appliance non-identifiable. The increase in $\epsilon_r$, associated to an appliance $r$, is expected to reduce the percentage of identifiable appliance signature. Fig 4a reports the percentage of identifiable appliance signature for each individual appliance in the smart meter data for a single day when, $\epsilon_r = 0 \forall r$. An $\epsilon_r = 0$ is the minimum privacy setting and means that the appliance signature should be left un-altered. Fig 4b, shows a different setting. Each individual appliance is assigned a different privacy setting using different $\epsilon$ values. Fridge $\epsilon_1 = 0.5$, Dryer $\epsilon_2 = 0.8$, Kettle $\epsilon_3 = 0.7$, Washing Machine $\epsilon_4 = 0.2$, PC $\epsilon_5 = 1$ and Freezer $\epsilon_6 = 0.3$. It is observed that the percentage participation of each individual appliance reduces according to its own epsilon value. For example, the freezer whose $\epsilon_6$ has increased from 0 to 0.3 has reported a participation decrease from 30% to 24%. The decrease in the participation of all the appliances has resulted in the increase of the Other category, which refers to the consumption reading that does not fall into any of the appliance signatures.

We further elaborate on the change in percentage contribution of each appliance signature when the privacy setting $\epsilon_r$ for only a single appliance is altered at a time, keeping the rest unaltered. Table 2 reports the results when $\epsilon$ for a single appliance is changed from zero to 1 in steps of 0.25 and the rest are kept zero. It can be observed from Table 2 that, as the privacy for a particular appliance increases with an increasing value of $\epsilon$, the percentage contribution, of that particular appliance decreases.

The original data set shows that, the overall power consumption of the household is divided into three phases. The Fridge and Kettle work on one phase, the dryer works separately on one phase and the Washing Machine along
with the PC and Freezer work on the remaining phase. In a scenario when the user is willing to hide only a single appliance from the power consumption pattern, an efficient privacy preserving algorithm would be expected to identify the appliance correctly even if it is working concurrently with multiple other devices and hide the signature accordingly. This would, as a result, affect the percentage contribution of the devices that are working on the same time instance as the concerned device. This phenomenon is also visible from Table 2 where, by changing the privacy setting of the Fridge, the percentage contribution of the Kettle is also being affected and likewise, by changing the privacy setting of the Freezer, the contribution of the PC and Washing Machine are also affected, hence highlighting that the proposed approach correctly identifies appliances that are working concurrently.

We further validate the performance of the proposed approach using information theory principles. In particular, we observe the change in mutual information between the original smart meter reading and the smart meter reading after it is scattered using the proposed approach. Mutual information quantifies the amount of information two data sets share and thus how much does one data set reveal about the other [41]. We use mutual information between the original and the scattered data set as a measure of utility. Fig 5 reports the change in mutual information when \( \epsilon \), associated to a particular appliance is increased from zero to one thus increasing the privacy of the appliance. An increase in privacy should result in a decrease in mutual information and thus a decrease in utility of the data. However, the question which arises is, which appliances have the most effect on the overall privacy. Fig 5 reports that the larger the contribution of an appliance reported in Fig 4a, the greater the decrease in mutual information thus a greater decrease in the utility of the data.

![Figure 5: Change in mutual information between the Original data and the smart meter data scattered using the proposed approach for increasing \( \epsilon \).](image)

Mutual information can also be used to assess privacy where privacy is a measure of the percentage contribution of each individual appliance. An increase in privacy in the considered environment implies that the percentage contribution of each individual appliance signature should reduce. We test this by reporting the change in mutual information between the smart meter data scattered using the proposed approach and the plug data available for each appliance. Fig 6 reports that, as the privacy setting is increased for each individual appliance, the mutual information between the plug data and the scattered data decreases thus showing an increase in privacy. The largest rate of decrease is reported by the Freezer which has the highest contribution to the overall consumption.

![Figure 6: Change in mutual information between the plug data and the smart meter data scattered using the proposed approach for increasing \( \epsilon \).](image)

We complement our claims using another set of simulations. In these simulations, we test the privacy preserving capability of the proposed approach against two well-known Non-Intrusive Load Monitoring (NILM) algorithms, namely: The Weiss’ algorithm and the Baranski’s algorithm. The time-series data generated using the proposed approach is analysed by these NILM algorithms and the number of appliances being identified reported. The larger the number of successfully identified appliances, the lower the level of privacy. Therefore, in ideal case, maximum privacy would be assured when no appliance is being detected.

NILM algorithms are used to identify appliances in a household and these algorithms can be broadly categorized into three classes: unsupervised, semi-supervised and supervised. An unsupervised algorithm only requires a household’s aggregate electricity consumption data, a semi-supervised model utilizes generic appliance models thus avoiding the requirement of intrusively installing sensors or the use of other training methods, whereas, a supervised approach uses the plug level data along with the smart meter reading to train and then to identify the existence of an appliance [42]. For our experiments, we use two NILM algorithms, a supervised algorithm referred to as the Weiss’ algorithm [15] and an unsupervised algorithm referred to as the Baranski’s algorithm [42]. The selection of the Weiss’ algorithm is made with the understanding that an approach working well for a supervised algorithm will naturally be better in performance when tested against semi-supervised or unsupervised algorithms. The Weiss’ Algorithm uses both the smart meter and the plug data available for each household to identify the real and reactive power of each appliance being used in the household. It extracts switching events from the household’s aggregate electricity consumption and assigns each event to the appliance with the best match in a signature database. On the other hand, the Baranski’s algorithm...
Table 3: Energy consumption data obtained using the Weiss’ algorithm

<table>
<thead>
<tr>
<th>Appliance</th>
<th>$e_r = 0.05r$</th>
<th>$e_r = 0.25r$</th>
<th>$e_r = 0.55r$</th>
<th>$e_r = 0.755r$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Power</td>
<td>Reactive Power</td>
<td>True Power</td>
<td>Reactive Power</td>
</tr>
<tr>
<td>Freezer</td>
<td>-30.25</td>
<td>-29.91</td>
<td>-32.25</td>
<td>-37.38</td>
</tr>
<tr>
<td>Freezer</td>
<td>29.52</td>
<td>27.61</td>
<td>178.36</td>
<td>NaN</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>-2100.37</td>
<td>3.65</td>
<td>-731.76</td>
<td>-63.36</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>2195.10</td>
<td>10.21</td>
<td>735.74</td>
<td>57.49</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>793.83</td>
<td>306.61</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PC</td>
<td>64.55</td>
<td>-14.26</td>
<td>32.08</td>
<td>-2.53</td>
</tr>
<tr>
<td>PC</td>
<td>92.36</td>
<td>68.29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fridge</td>
<td>-24.12</td>
<td>-0.21</td>
<td>-151.02</td>
<td>Inf</td>
</tr>
<tr>
<td>Fridge</td>
<td>24.05</td>
<td>0.11</td>
<td>152.52</td>
<td>NaN</td>
</tr>
<tr>
<td>Fridge</td>
<td>61.15</td>
<td>-12.66</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>-1824.00</td>
<td>28.61</td>
<td>-1257.38</td>
<td>-22.50</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>-1846.49</td>
<td>-43.21</td>
<td>-635.30</td>
<td>13.73</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>1872.67</td>
<td>73.17</td>
<td>1753.02</td>
<td>64.32</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>1850.34</td>
<td>-22.55</td>
<td>1108.42</td>
<td>271.07</td>
</tr>
<tr>
<td>Dryer</td>
<td>-394.89</td>
<td>-181.71</td>
<td>318.42</td>
<td>-7.27</td>
</tr>
<tr>
<td>Dryer</td>
<td>544.54</td>
<td>172.12</td>
<td>360.20</td>
<td>255.35</td>
</tr>
<tr>
<td>Dryer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: Energy consumption data obtained using the Weiss’ algorithm

<table>
<thead>
<tr>
<th>Appliance</th>
<th>Proposed Approach $e_r = 1/r$</th>
<th>ElecPrivacy $[13]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Power</td>
<td>Reactive Power</td>
</tr>
<tr>
<td>Freezer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Freezer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PC</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PC</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fridge</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fridge</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fridge</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Water Kettle</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dryer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dryer</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The selection of this approach is due to its most relevance to the appliance as the Weiss’ algorithm does not identify all the appliances, if required by the user, with the hiding being confirmed when Weiss’ and/or the Baranski’s algorithm are not able to detect the appliances in a household. Table 3 reports the results when smart meter data is scattered using the proposed approach and the resulting scattered data is fed into the Weiss’s algorithm to identify appliances in the household. In this set of experiments, $e_r$ is increased from 0 to 0.75 in steps of 0.25. The Weiss’s algorithm is fed with the original plug data along with the altered smart meter data and then the appliance utilization is identified. Table 3 shows the change in true power and reactive power at On and OFF switching events, where a positive true power refers to an On switching event and a negative true power refers to an OFF switching event. It is evident that as $e_r$ increases from 0 to 0.75, the magnitude of true and reactive powers identified for each individual appliance decreases. This shows that with an increase in $e_r$ fewer number of switching events are classified for each individual appliance thus ensuring an increase in privacy. A reactive power of $Inf$ and $NaN$ is a result of false detection of an appliance. In the ECO data set, the smart meter power readings are accompanied by the corresponding changing current and voltage values and when the current and voltage curves do not support the power curve, the Weiss’ algorithm reports a $NaN$ or $Inf$ value. Thus it is clear from Table 3 that the proposed approach works considerably well despite a low accuracy and precision value reported in Table 1. We also compare the proposed approach with $e_r = 1$ against the existing ElecPrivacy approach and report the observed results in Table 4. It is observed that the proposed approach performs well in hiding an appliance as the Weiss’ algorithm does not identify any appliance whereas, it identifies most of the appliances when the existing ElecPrivacy approach is used. The main reason for this identification of appliances is that they divide the battery energy into equal number of slots, where is unsupervised and thus operates without having information about the appliances existing in a household. It identifies recurring electricity consumption events in the aggregate electricity consumption of a household and clusters these events assuming that events in the same cluster belong to the same appliance. In these experiments, we also compare the proposed approach with an existing approach referred to as the ElecPrivacy [13] approach. The selection of this approach is due to its most relevance to the proposed approach as the authors therein identify the importance of individual appliances in a household. They divide the storage capacity of the battery source into equal sized slots where the number of slots is equal to the number of appliances in the household and use a unique charge and discharge pattern to hide all the appliances together.

We evaluate the performance of the proposed approach against the existing ElecPrivacy approach by testing the generated scrambled time-series data using the Weiss’ and the Baranski’s algorithm. An algorithm would be considered efficient in ensuring privacy if it is able to hide all the appliances, if required by the user, with the hiding being confirmed when Weiss’ and/or the Baranski’s algorithm are not able to detect the appliances in a household. Table 3 reports the results when smart meter data is scattered using the proposed approach and the resulting scattered data is fed into the Weiss’s algorithm to identify appliances in the household. In this set of experiments, $e_r$ is increased from 0 to 0.75 in steps of 0.25. The Weiss’s algorithm is fed with the original plug data along with the altered smart meter data and then the appliance utilization is identified. Table 3 shows the change in true power and reactive power at On and OFF switching events, where a positive true power refers to an On switching event and a negative true power refers to an OFF switching event. It is evident that as $e_r$ increases from 0 to 0.75, the magnitude of true and reactive powers identified for each individual appliance decreases. This shows that with an increase in $e_r$ fewer number of switching events are classified for each individual appliance thus ensuring an increase in privacy. A reactive power of $Inf$ and $NaN$ is a result of false detection of an appliance. In the ECO data set, the smart meter power readings are accompanied by the corresponding changing current and voltage values and when the current and voltage curves do not support the power curve, the Weiss’ algorithm reports a $NaN$ or $Inf$ value. Thus it is clear from Table 3 that the proposed approach works considerably well despite a low accuracy and precision value reported in Table 1. We also compare the proposed approach with $e_r = 1$ against the existing ElecPrivacy approach and report the observed results in Table 4. It is observed that the proposed approach performs well in hiding an appliance as the Weiss’ algorithm does not identify any appliance whereas, it identifies most of the appliances when the existing ElecPrivacy approach is used. The main reason for this identification of appliances is that they divide the battery energy into equal number of slots, where
this number is decided by the number of appliances in a household. Each appliance uses the available slot to hide its signature and the battery is charged when no appliance is consuming energy. In the considered ECO data set there are 6 appliances and very few intervals during which the battery can be charged using their defined approach. On the contrary, the proposed approach uses a novel battery charging pattern which ensures that the battery is charged at all possible intervals using a different amount of energy, hence enabling us to add noise into the smart meter reading.

We further strengthen our privacy claim by testing the proposed approach against the unsupervised Baranski’s algorithm. Similar to the previous experiment, the Baranski’s algorithm is fed as input the time-series data scrambled using the proposed approach with varying $\epsilon_r$ from 0 to 1 in steps of 0.25 and comparing the obtained results with the ElecPrivacy approach. Privacy is measured by the number of appliances that are successfully identified by the Baranski’s algorithm, where a higher number of identified appliances is interpreted as a lower level of privacy.

The Baranski’s algorithm takes as input the aggregate 1Hz electricity consumption data for the household. We set the resulting number of clusters to 20 with the baseline condition that each appliance can only have two states with a maximum ON state of 60 minutes. Table 5 reports the results from the experiment. Each cluster shows the set of switching events that report a similar change in the electricity consumption pattern with the column size representing the number of events in a cluster. To enhance the usability and understandability of the obtained results, time stamps in the plug level information are used to assign each event to an appliance. The percentage of each assigned event in a cluster is reported in the % column and it is calculated using the proportion of events assigned to the appliance divided by the overall number of events in the cluster. Column App.1 and App.2 represent the appliances that have the highest and second highest number of assigned events in a cluster respectively. It is noticeable that, when no privacy is introduced with $\epsilon_r = 0 \forall r$, Cluster 4 and Cluster 6 exclusively contain stop and start event of the Freezer respectively. Whereas, Cluster 1 reports a switching event due to both the Washing Machine and the Freezer. Table 5 highlights that, as $\epsilon_r \forall r$ increases from 0 to 1, the overall size of clusters increase resulting in a decrease in participation of individual appliances. On the same time, the magnitude of change in power $\Delta P$ also decreases with increasing $\epsilon_r$. These changes are merely a result of the increase in noise in the generated time-series data. The addition of noise divides a single high power spike into multiple small steps thus generating a large number of clusters with smaller $\Delta P$ values. In the considered household, the Fridge and the Freezer consume the smallest amount of RMS power, this results in the identification of mostly these devices when more noise is added into the generated time-series data. In a household, a Fridge and a Freezer run continuously for the whole day and they show no correlation with the presence or absence of an individual, therefore the increase in number of clustered reporting Fridge and/or Freezer is referred to as the success of the proposed approach. On the contrary, when the same Baranski’s algorithm is used to identify appliances from the time-series data generated using the ElecPrivacy approach, table 5 reports that, most of the appliances are still easily identifiable with only a slight change in the magnitude of $\Delta P$.

![Figure 7: Smart meter reading of a household on 7/06/2012 for comparison between the a) original reading b) after scrambling using the proposed approach c) after scrambling using the existing ElecPrivacy approach.](image)

In the final set of results, we demonstrate the beneficial effect of noise addition to the proposed approach. Fig 7 shows the resultant smart meter data for a randomly selected date (7/06/2012) when the proposed approach and the ElecPrivacy approach are used. It also presents the original smart meter data for reference. In this set of experiments, we use a harvesting interval of 5 minutes (300 sec). This creates the need for any privacy ensuring algorithm to ensure that the average power consumption for each 300 sec interval remains the same despite the necessary scattering. Fig 7b shows that the proposed approach maintains the same energy consumption pattern as that reported in Fig 7a whereas, ElecPrivacy in Fig 7c reports a different pattern. At the initial stages from time values $2.2 \times 10^4$ and $2.4 \times 10^4$ there is an extra peak. This does
Table 5: Event clusters provided by Baranski’s algorithm

<table>
<thead>
<tr>
<th>Cluster</th>
<th>$\Delta P$</th>
<th>$\epsilon_r = 0.05\epsilon$</th>
<th>$\epsilon_r = 0.25\epsilon$</th>
<th>$\epsilon_r = 0.75\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>-42</td>
<td>15% Freezer 1% Washing Machine</td>
<td>3% 54904% Washing Machine 1% Freezer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_2$</td>
<td>54</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_3$</td>
<td>35</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_4$</td>
<td>40</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_5$</td>
<td>34</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_6$</td>
<td>47</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_7$</td>
<td>43</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_8$</td>
<td>43</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_9$</td>
<td>36</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_10$</td>
<td>37</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_11$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_12$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_13$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_14$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_15$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_16$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_17$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_18$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_19$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
<tr>
<td>$c_20$</td>
<td>28</td>
<td>1% Freezer 1% Washing Machine</td>
<td>1% 33040% Washing Machine 1% Dryer</td>
<td>1% Washing Machine 1% Freezer</td>
</tr>
</tbody>
</table>

$\Delta P$: Average pressure change; $\epsilon_r$: Relative error; $\epsilon = 0.05\epsilon$: Cluster with low relative error; $\epsilon = 0.25\epsilon$: Cluster with moderate relative error; $\epsilon = 0.75\epsilon$: Cluster with high relative error.
hide an appliance that might be running on the particular instance but it changes the average consumption per harvesting interval thus forcing the consumer to pay more due to the altered smart meter reading. Moreover, a pattern similar to the original meter is reported in the interval from $2.4 \times 10^4$ and $2.8 \times 10^4$ as a result of the battery not having sufficient energy to alter the smart meter reading thus revealing all the appliances.

6. Conclusion

In this work, we highlight the relationship between appliance specific privacy of an individual and the utility of a smart meter generated data. We proposed a new approach that allows the user to control the amount of information that is being shared regarding each individual appliance. This allows the user to trade between the level of user data privacy in exchange to the utility of the resultant smart meter data. Simulations were conducted to test the proposed approach with respect to the amount of privacy achieved and the resulting smart meter data utility. Results indicate the proposed approach is successful in controlling the privacy of each individual appliance and the utility of the resulting smart meter data. Appliances with higher power consumption are also found to have a larger effect on the total privacy achieved. The obtained results also indicate an improvement in performance of the proposed approach when compared with the existing ElecPrivacy approach. This comparison was conducted by feeding the resulting smart meter data into the Weiss’ and the Baranski’s algorithm which highlight that, the traces generated by the proposed approach are harder to identify. The latter demonstrates its effectiveness. In the future, we aim at adopting a more analytical approach to the considered problem by attempting a direct or sub-optimal solution to the posed optimization problem.

References


[35] M. Chessa, J. Grossklags, F. Loiseau, A game-theoretic study on non-monetary incentives in data analytics projects with privacy