ElecPrivacy: Evaluating the Privacy Protection of Electricity Management Algorithms

Georgios Kalogridis, Rafael Cepeda, *Senior Member, IEEE*, Stojan Z. Denic, *Member, IEEE*, Tim Lewis, and Costas Efthymiou, *Member, IEEE*

***Abstract*—The data collected by a home smart meter can potentially reveal sensitive private information about the home resident(s). In this paper, we study how home energy resources can be used to protect the privacy of the collected data. In particular we: a) introduce a power mixing algorithm to selectively protect a set of consumption events; b) develop a range of different privacy protection metrics; c) analyze real smart metering data sampled twice a minute over a period of 13 days; and d) evaluate the protection offered by different power mixing algorithms. Major factors which determine the efficiency of the proposed power mixing algorithms are identified, such as battery capacity and power, and user preferences for privacy-based allocations of battery energy quotas.**

***Index Terms*—Smart metering privacy, energy management, rechargeable batteries, power routing.**

I. Introduction

SMART GRIDS (SGs) are an emerging engineering effort to reform the world's electrical grids [1]; they interconnect their components with a two-way communications network to support real-time optimizations such as load shedding/management, distributed energy storage (e.g., in electric vehicles), and distributed energy generation (e.g., from renewable resources). Smart meters have a pivotal role in this infrastructure: they can measure and communicate detailed real-time energy usage technical data to the grid (e.g., utility providers), and facilitate remote power monitoring and control.

Smart grid privacy encompasses the privacy of information obtained from smart metering data, which may be systematically collated and analyzed. Privacy concerns may be demonstrated with the use of *nonintrusive appliance load monitors* (NALM), which analyze energy signatures to track appliance usage patterns [2]. There is a rich and ongoing line of research in the construction and upkeep of NALM algorithms, providing means to identify appliance usage even when multiple household load signatures are aggregated [3]. Quinn [4] argues that frequently collected metering data, e.g., at 15 minute intervals, may provide a window into the activities within homes, exposing a wealth of private activities to anyone with access to energy usage information.

The need for SG privacy is recognized by many standardization bodies, such as the National Institute of Standards and Technology (NIST) in the United States [5]. Much recently, NIST published a set of guidelines focusing on the development of privacy policies and practices [6]. These may be based on standard *privacy principles* such as notice, choice and consent. In Europe, the Steering Committee has set up a SG Expert Group to identify the appropriate regulatory scenario and recommendations for data handling, security, and consumer protection [7].

While robust privacy policies are needed to regulate usage of SG data, we consider that certain privacy aspects of smart metering data could be best protected by design. In this direction, anonymity services may help to protect privacy. For example, metering data can be aggregated and encrypted so that an individual's information is anonymized to roughly the scale of a city block [4]. An elaborate smart metering anonymity protocol has been proposed in [8] where it is assumed that metering data can be separated into low frequency attributable data (e.g., data used for billing) and high frequency anonymous technical data (e.g., data used for demand-side management). In such schemes the offered privacy will still depend on trust relationships and data policies that govern the involved parties.

In [9] the authors introduce an alternative protection scheme where the energy flow within a home is controlled by running a portion of a consumption demand off a rechargeable battery, rather than directly off the grid. That is, energy flow can be managed in a manner advantageous to customer privacy if characteristic load signatures can be masked in a way that makes it harder to detect appliance usage patterns. This paper focuses on this system, which we call "ElecPrivacy," and work is extended in a number of ways as follows. We introduce a selective privacy protection algorithm using prediction and customized privacy preferences, extend the protection metrics from [9], and use smart metering data collected from individual home appliances to evaluate the protection offered by ElecPrivacy.

The rest of this paper is organized as follows. The ElecPrivacy system and experimental methodology are discussed in Section II. The power consumption prediction model and privacy protection algorithm are given in Section III. The privacy protection metrics and data evaluations are analyzed in Section IV, and conclusions are drawn in Section V.

II. ElecPrivacy System Overview

A. ElecPrivacy Concept and System Implementation

In a smart grid system where power generation is episodic, invariable and distributed (e.g., originating from renewables), energy storage is one of the key underpinnings [10]. The ElecPrivacy system shares this vision and it assumes the existence of a) an energy storage facility, such as electric vehicle (EV) bat-

Manuscript received October 16, 2010; revised April 15, 2011; accepted June 05, 2011. Date of publication August 30, 2011; date of current version November 23, 2011. Paper no. TSG-00187-2010.

The authors are with the Telecommunications Research Laboratory, Toshiba Research Europe Limited, Bristol, BS1 4ND, U.K. (e-mail: george@toshiba-trel.com; rafael@toshiba-trel.com; stojan.denic@toshiba-trel.com; timl@toshiba-trel.com; costas@toshiba-trel.com).

Digital Object Identifier 10.1109/TSG.2011.2160975

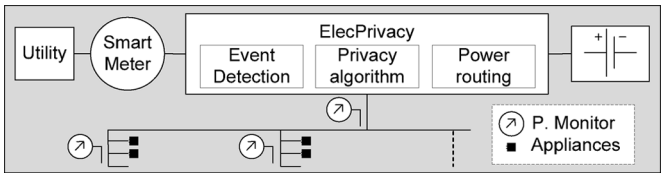


Fig. 1. ElecPrivacy system overview.

tery, and b) an "electrical power routing" mechanism, where this term is taken to mean the selective control and power mixing of a number of electricity sources to cover consumption demands [11]. For example, ElecPrivacy may be implemented within a "charge grid system" [12] that may use a rechargeable battery and a bidirectional inverter to optimize the flow and storage of electricity.

The ElecPrivacy algorithm differs from other energy management algorithms in the manner with which the battery charges or discharges. In this sense, ElecPrivacy can be considered as a privacy-driven power mixing algorithm, where the power mixing is conducted not in response to energy shortage/surplus or cost, but (primarily) for privacy purposes. For example, suppose that we wish to use battery energy to hide the demand of an appliance. In this case, the use of battery energy will shed the composite peak. This might help reduce energy costs (e.g., during a high tariff period), but it might as well reduce the efficiency of energy pricing arbitrage (e.g., during a low tariff period) [9].

An overview of the ElecPrivacy system can be seen in Fig. 1, comprising the following subsystems.

- *Metering mechanism*: used to obtain a set of electricity measurements that correspond to a set of consumption events. This mechanism may be implemented with a set of power monitors attached to different devices and/or a load signature analysis method used to extract such information from the composite metering data.
- *Event detection*: analyzes event information in order to detect an occurring, or predict an imminent, "privacy threat." For example, this may be a power trigger generated by a particular event, such as a change in power consumption (e.g., appliance switch-on/off event).
- *Privacy protection algorithm*: configures power routing in order to mask a detected consumption event.
- *Power routing*: mixes a private (i.e., nonutility) energy resource (e.g., rechargeable battery) with utility energy to meet appliance demands.

B. Experimental Setup

For the purposes of this paper an ElecPrivacy metering system was set up in a studio apartment and operated for 13 days, with the data gathered being representative of an apartment of this type. Detailed energy usage patterns were captured by installing a total of 10 ZigBee electricity monitors, each one measuring the consumption of a set of appliances S_i within the apartment, where i indicates the used meter number, as seen in Table I. These electricity monitors [13] can be used both "inline" (plugged into an electrical socket) or as passive monitors with a current core transducer probe clipped over a cable, which also allowed measurement of the whole apartment's consumption.

TABLE I
POWER METERS USED IN THE EXPERIMENTS AND AMOUNT OF AVERAGE ENERGY USAGE IN PERCENTAGE

Meter No.	Electrical appliance	Power (%)
1	Home laptop, Heater	3.13 %
2	Raclette, Battery charge Vacuum cleaner	11.92 %
3	Kettle, Toaster, Hair clipper	3.39 %
4	Bread maker, Milk maker, Blender	4.01 %
5	Laptop (meters)	10.91 %
6	Cooker	2.66 %
7	Washer	2.56 %
8	Fridge	23.60 %
9	Oven	1.48 %
–	Others	36.32 %
10	General consumption (Average daily consumption: 2930.2 Wh)	100.00 %

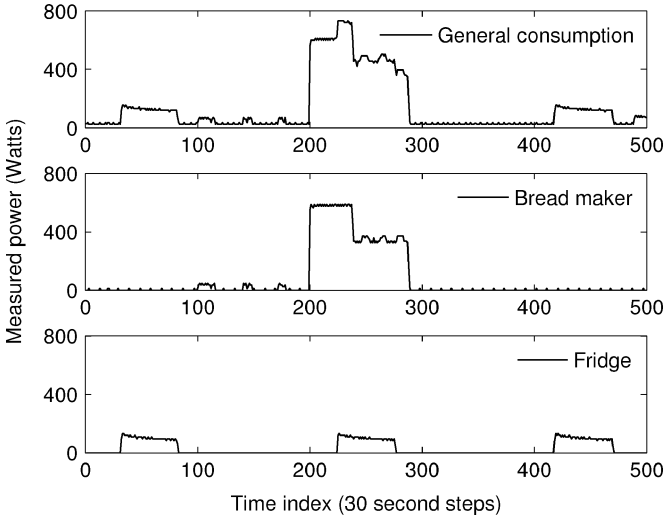


Fig. 2. Example of measured data.

The electricity monitors were configured to take measurements every 30 s and were all synchronized. The sampling interval was limited to 30 s to allow reliable periodic data gathering from all 10 wireless monitors. We suppose that this interval is shorter than the one utility companies may use to gather smart metering data in the future (e.g., 15 min [4]). The objective of our measurement campaign was to obtain data in as granular a way as possible. By using such high frequency data the ElecPrivacy system can more effectively respond to shield against potential privacy threats using the methods discussed in this paper. A characteristic sample window of individual appliance and aggregate power loads is shown in Fig. 2.

It is noted that this paper only considers the average power usage data (over each sampling interval) derived from consecutive energy measurements. We prefer such average power values over instantaneous power samples, as the latter do not capture potential power draw fluctuations occurring between samples. More precisely, we define a power load signature (PLS) to be a trace of time-stamped average power loads $p(t)$ derived from a trace of (cumulative) energy values $e(t)$ metered at intervals Δt, $p(t) = (e(t) - e(t - \Delta t))/(\Delta t)$. A home PLS, or compound PLS (CPLS), is the sum of all home appliance loads.

We also note that the logging precision of the used electricity monitors is 0.1 Wh, which corresponds to the amount of energy used by a 12 W load over a 30 s period. Thus, the obtained average power consumption figures are "quantized" into multiples of 12 W, yielding a maximum error of ± 6 W in each of the values of $p(t)$.

C. Assumptions and Objectives

There are many different definitions of smart grid privacy; in this paper we consider that the undetectability of individual appliance events offers "privacy of personal behavior," i.e., "the right of individuals to keep any knowledge of their activities, and their choices, from being shared with others" [6]. For example, the knowledge of an appliance operation may imply home occupancy or vacancy and lifestyle information such as sleeping times. Such privacy threat analysis is out of this paper's scope. Instead, the aim is to hide appliance operation events and measure the level of the offered protection. We suggest that PLS privacy comprises information of consumption events, which we wish to hide, obfuscate, or emulate (in real time) so that they cannot be identified (i.e., extracted from the PLS). Adapting the definition of "undetectability" from [14], we consider that privacy is protected when, given a PLS, we cannot sufficiently distinguish whether a particular appliance event exists or not.

Privacy protection may be discussed in the context of anonymity: the property of hiding the identity of a user associated with a message (rather than hiding the message itself) [15]. Anonymity can be analyzed using information theoretic metrics, for which there is a line of active research [16], [17]. The PLS "undetectability" problem differs in that we wish to hide some PLS characteristics rather than hide the user; however, an analogy may be drawn between the "user" and PLS [9].

In this paper we consider the somewhat strong privacy threat scenario where the adversary: a) knows the energy load profiles for all appliances, and b) can accurately decompose (e.g., by means of edge detection [2]) a (recorded) CPLS into a set of PLSs of distinct appliances. That is, the adversary can distinguish appliance events and it can correlate them with their energy profiles. The objective is thus to modify an appliance PLS so as to reduce the likelihood the original power trace can be derived from the modified power trace. This extends the work in [9] where it was proposed to modify CPLS rather than appliance PLSs.

In this direction, we introduce in Section III-C a privacy protection algorithm that selectively protects a set of home appliances. The objective is to distribute energy storage resources in a manner that offers different levels of protection for different appliances. That is, the allocation of privacy resources can be personalized. We also consider a simple privacy threat example where the probability of the operation of an appliance yields a degree of probability of home occupancy, as described in Table II. Based on this, we assume that the fridge PLS does not need to be protected as it is unlikely to provide any useful indication of occupancy. (Note here that in our measurements the fridge power load pattern does not vary, regardless of related human activity, across weekends and weekdays.)

TABLE II
SPECTRUM OF HOME OCCUPANCY PROBABILITIES BASED ON APPLIANCE OPERATION

Probability	Device operation
Very high	Vacuum cleaner, Raclette, Kettle, Toaster Hair clipper, Blender
High	Laptop, Bread maker, Milk maker, Cooker, Oven
Medium	Washer
Unknown	Fridge, Battery charger

III. PRIVACY PROTECTION MECHANISM

A. Battery-Assisted Power Transformation

Suppose that a battery is discharged or recharged with a $p_B(t)$ average power over a Δt metering interval in order to "disguise" a given consumption load $p(t)$ (i.e., home load, excluding the battery). With the use of battery power mixing, the home power trace becomes $p' = p - p_B - p_L$, where $p_L(t)$ is the (average) battery power loss due to charging/discharging during $(t - \Delta t, t)$. The objective is to modify p_B in real time in order to obtain a p' that renders p undetectable, i.e., p cannot be determined given p'. This modification is bounded by the physical battery capabilities:

1) The battery has a finite capacity E_C (hence, it has to maintain its energy by recharging), i.e., $0 \leq \int_0^T p_B(t)dt \leq E_C$, for all $T \geq 0$ (assuming that for $t = 0$, the battery is fully charged).
2) The battery has a maximum discharge and recharge power of P_D and P_R, i.e., $-P_R \leq p_B(t) \leq P_D$, for all t.

Intuitively, the level of privacy protection should approach a maximum when $p'(t) = 0$, for $0 \leq t \leq T - \Delta t$ (the battery covers all the power demand), $p'(t) = \int_0^T p(t)dt/\Delta t$. for $T - \Delta t \leq t \leq T$ (battery fully recharges), $\Delta t \to 0$, and $T \to \infty$. However, this scheme would require $E_C \to \infty$. Alternatively, it should be equally secure to have $p'(t) = \int_0^T p(t)dt/T$, for $0 \leq t \leq T$ (the load p is transformed to a constant value representing its average): this moderation should use the battery more efficiently but, on the downside, it requires knowledge of all future consumption events.

The problem is how to maximize privacy protection, given a load p, battery capacity/power bounds (E_C, P_D, P_R), and no knowledge of the future. Suppose that a moderated home load p' is introduced by a transformation $\mathcal{G}$ on the (real-time) appliance load demand p such that $p' = \mathcal{G}p$. In order to evaluate (and optimize) $\mathcal{G}$ it is necessary to measure the level of protection it offers.

To evaluate $\mathcal{G}$, we first define what needs to be protected. Given a PLS, p, the difference between successive power measurements, $dp(t) = p(t) - p(t - \Delta t)$, represents a change (or not) of the state of some appliance as its power usage increases/decreases (or remains the same). We consider that any change (or no change) of appliance state, is "private information"; hence, we wish to measure the degree to which dp is detectable, given dp' after the application of $\mathcal{G}$.

Algorithm $\mathcal{G}$ should operate in real time. Thus, given the present values of $p(t)$ and $p'(t)$, $\mathcal{G}$ needs to predict energy demands and potential privacy threats, before computing the next $p'(t + \Delta t)$ modified power value. This formulates a problem of prediction as discussed in the following section.

B. Modeling Power Load Signature by Markov Chains

There are two reasons for finding appropriate mathematical representation of a CPLS as well as the PLSs of different appliances: it can be used for: 1) further theoretical analysis and 2) PLS prediction, which is an input for the privacy algorithm given in Section III-C.

We shall demonstrate how a simple Markov chain model can be applied in representing a PLS. This representation relies on the clustering analysis of the PLS discussed in Section IV-B. Denote the Markov chain representation by $X := \{X(k)\}_{k\geq 0},\ X(k) \in \{1,\ldots,N\}$. The Markov chain is completely defined by the transition probability matrix $\mathbf{A}(k)$ and the vector of state probabilities $\mathbf{P}(k) := [\Pr\{X(k)=1\},\ldots,\Pr\{X(k)=N\}]^T$. The transition probability matrix $\mathbf{A}$ contains conditional probabilities $a_{ij} := \Pr\{X(k+1)=i \,|\, X(k)=j\}$. Under the assumption that the PLS can be described by a stationary process, the evolution of the state probability vector is given by

$$\mathbf{P}(k+1) = \mathbf{A}\mathbf{P}(k), \tag{1}$$

when $\mathbf{P}(0)$ is known.

Since the PLS clustering analysis provides for the transition probability matrix $\mathbf{A}$ between clusters (i.e., Markov states), the prediction of the PLS value can be carried out via its Markov chain representation. In the prediction, two possible cases can be considered: a fully observable case and partially observable case. The fully observable case means that the state of the Markov chain $X(k)$ can be measured accurately, while the latter means that the Markov chain observation X is corrupted by noise. The noisy observation is denoted by $Y := \{Y(k)\}_{k\geq 1}$. The theory of hidden Markov models gives the following recursive filter which can be used for the one-step prediction [18]

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{\Gamma}\mathbf{Q}(k), \tag{2}$$

where $\mathbf{Q}(k)$ is the so-called unnormalized conditional state probability vector, $\mathbf{\Gamma}$ is a diagonal matrix having a vector $N[\Pr\{Y(k+1) \,|\, X(k)=1\},\ldots,\Pr\{Y(k+1) \,|\, X(k)=N\}]^T$ on the main diagonal, and $\mathbf{Q}(0)=\mathbf{P}(0)$. The conditional probability (conditioned on the past observations) of being in the state i at the time instant k is determined by

$$P_i(k) = \frac{Q_i(k)}{\sum_{i=1}^{N} Q_i(k)}, \tag{3}$$

where $P_i(k)$ and $Q_i(k)$ are the ith entries of the vectors $\mathbf{P}(k)$ and $\mathbf{Q}(k)$, respectively. For the time instant $k+1$, the prediction of the state $\hat{X}_{k+1}$ is obtained by using the maximum likelihood (ML) principle, by choosing the state i having maximum probability of occurrence.

Fig. 3 shows the performance of the one-step prediction algorithm for a measured CPLS. It can be observed that the estimator gives an accurate prediction of the measured signal confirming that the Markov chain model is a good description of the PLS. A similar conclusion is true for the prediction of particular appliances.

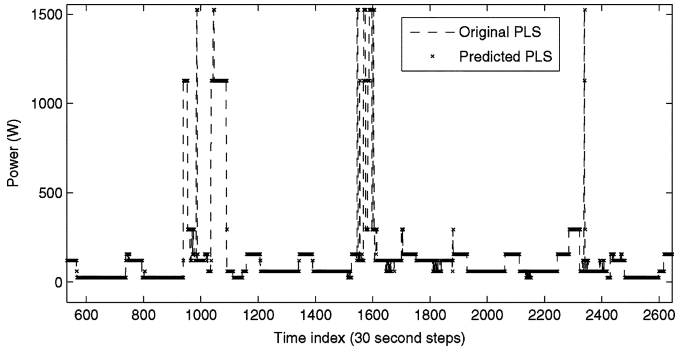


Fig. 3. Original CPLS signal and its one-step prediction.

C. Selective Privacy Protection Algorithm

We now extend the "best-effort" privacy algorithm in [9], which simulates water-filling [19], by selectively allocating logical battery resources to different appliances.

First, we allocate a set of logical battery energy quotas, $E_i(0)$, to a corresponding set of appliances S_i we wish to hide, where $E_i(0)$ denotes the available logical battery energy level for S_i at step $t=0$. We then choose $E_i(0)$ so that $\sum_i E_i(0) = E_C$. Intuitively, we compute $E_i(0)$ by normalizing the energy usage percentages in Table I after the removal of the "Fridge" and "Others." However, different, customized energy quotas may be chosen to suit different privacy requirements, such as the ones related to the event of home occupancy highlighted in Table II. For example, a larger $E_i(0)$ may increase the privacy of S_i at the expense of other appliances' privacy.

At each time slot t_1 a prediction of the subsequent power value $p_i(t_2)$ is made (as discussed in Section III-B), where i is an index for S_i. The modified $p_i'(t_2)$ resists to the possible degree against the predicted S_i operational state changes, according to the "legacy" best-effort algorithm. That is, the metered (modified) load p_i' remains constant as long as $E_i(0)$ can absorb p_i power demands. The algorithm will force an S_i logical battery charge level $E_i(t)$ at time t to either discharge or recharge, if the predicted $p_i(t_2)$ is either larger or smaller (respectively) than the current $p_i'(t_1)$.

Accounting for the trace of logical $E_i(t)$ requires us to calculate a corresponding trace of logical average battery charge/discharge power $p_{B,i}(t)$. Intuitively, an appliance may potentially use the maximum charge or discharge battery power (in contrast with $E_i(t)$ which can be no more than a percentage of E_C). We assume that there are no battery power losses, i.e., $p_L = 0$. At each time slot t_1 the algorithm determines how the available (average) battery power may be shared at the subsequent time slot t_2 as follows. Each appliance i is expected to require battery power of $\beta_i = p_i''(t_2) - p_i'(t_1)$, where $p_i''(t_2)$ is the predicted appliance power at t_2. Suppose $\beta_R = \sum_{\beta_i<0} |\beta_i|$ and $\beta_D = \sum_{\beta_i>0} |\beta_i|$. The required battery power is estimated as $\hat{p}_B(t_2) = \beta_D - \beta_R$. If the battery can provide the required expected power, $-P_R \leq \hat{p}_B(t_2) \leq P_D$, then each appliance can use battery power $p_{B,i}(t_2) = \beta_i$. In a different case, if $\hat{p}_B(t_2) > P_D$ then the allocated power for an appliance that requires power from the battery (discharge) is reduced as follows: if $\beta_i > 0$, then $p_{B,i}(t_2) = \beta_i(P_D + \beta_R)/\beta_D$, otherwise $p_{B,i}(t_2) = \beta_i$. In

TABLE III
SELECTIVE APPLIANCE PRIVACY PROTECTION ALGORITHM $\mathcal{G}$

```
∘ Calculate remaining stored energy level for each appliance:
  E_i(t) = E_i(0) − e_i(t − Δt) + e'_i(t − Δt)
∘ Predict required battery power for each appliance β_i
∘ Adjust the battery power allocated to each appliance p_{B,i}(t)
for All appliances i do
  if β_i > 0 (discharging case) then
    if Enough share E_i(t) to supply β_i > 0 for Δt then
      ∘ Calculate allocated battery energy ΔE_i(t) = β_i
    else
      ∘ Use remaining allocated battery energy ΔE_i(t) = E_i(t)
    end if
  end if
  if β_i < 0 (charging case) then
    if Enough battery 'emptiness' to absorb |β_i| for Δt then
      ∘ Calculate allocated battery charge ΔE_i(t) = β_i
    else
      ∘ Full battery share recharge ΔE_i(t) = E_i(0) − E_i(t)
    end if
  end if
end for
∘ Mix in (total) battery energy: e_B(t) = Σ_i ΔE_i(t).
```

another similar case, if $\hat{p}_B(t_2) < -P_R$ then the allocated power for an appliance that requires the battery to charge is reduced as follows: if $\beta_i < 0$, then $p_{B,i}(t_2) = \beta_i(P_R + \beta_D)/\beta_R$, otherwise $p_{B,i}(t_2) = \beta_i$. The powers $p_{B,i}(t_2)$ are further adjusted if the prediction is inaccurate. It trivially follows as a corollary that $-P_R \leq \sum_i p_{B,i}(t_2) \leq P_D$.

An outline of the above algorithm, which we call selective privacy protection algorithm and we denote as $\mathcal{G}_S$, is given in Table III.

IV. MEASURING PRIVACY PROTECTION

Anonymity can be quantified using information theoretic metrics such as *relative entropy* [16]. To measure the privacy protection offered by the real-time PLS transformation, we expand the privacy comparison techniques discussed in [9] as follows.

A. Variational Distance

To design a proper privacy algorithm $\mathcal{G}$, it is useful to establish certain tools for measuring the level of privacy offered by $\mathcal{G}$. In this paper, we propose one well known information theoretic quantity which can be used to compare two sources of information, the variational distance (VD) [19]. To employ the VD, we will assume that $p_A(t)$ and $p(t)$ can be modeled as stochastic processes with probability distributions P_1 and P_2. For two discrete random variables having probability mass functions P_1 and P_2, the VD $\mu(P_1 \,||\, P_2)$ is defined by [19]

$$\mu(P_1 \,||\, P_2) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P_1(x) - P_2(x)|. \tag{4}$$

The importance of the VD is that it quantifies the relation between two probability distributions P_1 and P_2. The VD is always positive, and for P_1 identical to P_2, it is zero.

Accordingly, the level of protection offered by a mapping $\mathcal{G}$ can be measured by the VD $\mu_{\mathcal{G}}(P_1 \,||\, P_2)$ such that the higher the level of protection offered by $\mathcal{G}$, the larger the VD.

One can think of other information theoretic quantities such as the entropy [19]. However, for the signals continuous in amplitude, the entropy can be negative, which is not a desirable feature. Another good candidate could be a mutual information, but we leave this for further research.

B. Cluster Similarity

As a second metric, this section describes a method introduced in [9] of building a similarity metric based on the concept of clustering the power levels available in the trace. Cluster techniques take a set of data with a distance metric and a parameter n denoting the desired number of clusters, and group the data points into n clusters that minimise the distance between points. The clusters are then specified by a set of n cluster *centers* around which the other values are grouped. Usually a multidimensional Euclidean metric is used, but in this analysis the data is one-dimensional so the distance metric is the absolute difference between power consumption values.

The issue of selection of n, is key to the clustering process. Here we use the method of silhouette maximization [20] to choose the best number of clusters for the set of data. The silhouette value of a point and a set of clusters is a measure of how well that point fits into its assigned cluster. By averaging over all points in the dataset (constructing the *silhouette average*), we can extract a measure of clustering quality in a manner that allows us to compare with other clusterings of different n. The set of silhouette averages over a range of n typically forms a single peak, increasing from $n = 2$ to a maximum and then rapidly tailing towards zero for large n. We can select the maximum, denoted the *silhouette optimal* n_{so} of the dataset, and use this to form the optimal cluster.

This value is dependent on the particular clustering method used, and for clustering heuristics that have a stochastic component will depend on a particular run of the algorithm. We use the "clara" method from [21] since it is deterministic and scales well to the relatively large datasets we use in this paper.

A further adaptation to the clustering process is dependent on the particular sparse structure of some of the power traces. For a trace that corresponds to a single or small number of devices, the values tend to be zero for long periods of time, interrupted by short periods of nonzero values. When clustered directly, these tend to produce a single zero-valued cluster, since the nonzero portion is statistically insignificant when compared to the zero portion. To properly analyze these traces we use the method of removing the zeros from the original trace, generating the clusters, then adding a zero value to the set of cluster centers. Note that we use the absolute values of the power level differences since this treats power increases and decreases identically, and makes the correspondence between two sets of cluster centers trivial.

We use the following method to compare how well the scrambled version hides the transitions of the original data.

1) Select the clustering size n by computing the silhouette optimal number of clusters from the original power sequence.
2) Generate n clusters from the original and classify each data point in the original sequence into clusters $1 \ldots n$, call this sequence A. The cluster centers are sorted in value.
3) Generate and classify the scrambled version into n clusters, sequence B, using the same number of clusters, but

TABLE IV
RESULTS FROM THE DIFFERENT PERFORMANCE METHODS

dp_i	VD		Clusters		Regression	
	B1	B2	B1	B2	B1	B2
	($\mu_{\mathcal{G}}$)	($\mu_{\mathcal{G}}$)	(n, % Ignored, % Hidden)	(n, % Ignored, % Hidden)	($1-R^2$)	($1-R^2$)
1	0.0739	0.0740	4, 99.97%, 61.54%	4, 99.97%, 60.00%	99.98%	99.95%
2	0.0330	0.0332	3, 99.95%, 47.08%	3, 99.98%, 66.67%	100.0%	100.0%
3	0.0030	0.0032	20, 99.92%, 96.55%	20, 99.98%, 88.89%	97.52%	99.90%
4	0.0495	0.0498	3, 99.92%, 35.48%	3, 99.97%, 45.45%	99.99%	94.33%
5	0.2183	0.2183	3, 99.99%, 60.00%	3, 99.99%, 66.67%	99.99%	99.99%
6	0.0251	0.0256	3, 99.90%, 22.22%	3, 99.96%, 33.33%	99.85%	100.0%
7	0.0061	0.0060	3, 99.97%, 33.33%	3, 99.98%, 14.29%	94.49%	99.99%
8[a]	0.0000	0.0000	7, 83.90%, 00.00%	7, 83.90%, 00.00%	00.00%	00.00%
9	0.0006	0.0006	4, 99.97%, 58.33%	4, 99.98%, 33.33%	91.54%	99.94%
10[b]	0.1091	0.1087	3, 68.03%, 01.36%	3, 68.03%, 01.14%	57.82%	62.78%
11[c]	0.3819	0.3848	3, 99.62%, 21.83%	3, 99.89%, 41.46%	89.88%	99.87%

VD: Variational distance;
[a] Unprotected socket;
[b] General consumption;
[c] Legacy algorithm [9].

allowing choice of fresh cluster centers appropriate to the dataset.

4) Cluster 1 will always correspond to a zero (or very low) power transition, so we strip points from cluster 1 from A and the corresponding values from sequence B, to give sequences A' and B'.
5) Compute the ratio of incorrectly classified transitions from sequence B, $(|A' \neq B'|)/(|A'|)$.

This yields a ratio that measures how effectively the original values have been hidden in the scrambled version, these are quoted as percentages for the "Hidden" values in Table IV.

C. Regression Analysis

As a third metric, we quantify privacy by combining a *cross correlation* and a *simple linear regression* procedure, as discussed in [9]. The idea is that the degree to which the modified power trace $dp'_i(t)$ "predicts" the original trace $dp_i(t)$ can be analyzed by *fitting* the traces and measuring their "distance" in the time domain. To this end, the *coefficient of determination* R^2 can be computed by estimating the error sum of squares $SS_E = \sum_t (dp_i(t) - dp'_i(t))^2$ and the regression sum of squares $SS_R = \sum_t (dp'_i(t) - \bar{dp}_i)$, where $(\bar{\cdot})$ represents the mean value. Then, it follows that

$$R^2 = 1 - \frac{SS_E}{SS_R + SS_E}, \quad 0 \leq R^2 \leq 1.$$

$R^2 = 1$ indicates that predictions are fully explained by the model; whereas $R^2 = 0$ indicates the opposite. In fact R^2 approaches zero when $SS_E \gg SS_R$ or when $SS_R \to 0$. The first case occurs when the noise increases, and the second occurs when $dp'(t)$ does not change much with respect to $\bar{dp}$. In either case, it is suggested that the closer the R^2 to zero, the higher the privacy protection level.

D. Experimental Evaluations

We evaluate the privacy protection of: a) the selective privacy protection algorithm introduced in Section III-C and b) the legacy water-filling algorithm introduced in [9] by operating the three given metrics on datasets obtained from the experimental system discussed in Section II-B. We test two different batteries B1, 1 KW/2 kWh, and B2, 2 KW/4 kWh. These could be projected by stacking up SCiB batteries [22]. The obtained results are summarized in Table IV.

We denote as p'_i the modified PLS obtained by operating the selective algorithm $\mathcal{G}_S$, where i is an index for the smart meters in Table I. We further denote as p'_{11} the modified CPLS obtained by operating the legacy transformation on the original CPLS, i.e., $p'_{11} = \mathcal{G}_L(p_{10})$, obtained from [9]. Note that p'_{10} denotes the modified CPLS obtained by operating $\mathcal{G}_S$ as described in Section III-C. It follows that

$$p'_{10} = \sum_{i \neq 8,10} p'_i + p_{10} - \sum_{i \neq 8,10} p_i, \tag{5}$$

where $p_{10} - \sum_{i \neq 8,10} p_i$ is the combined consumption of the fridge (p_8) and "Others" that are not protected by $\mathcal{G}_S$, as discussed in Section II-C.

Table IV indicates that p'_{11} yields better privacy protection as compared with p'_{10}. Similarly, in Fig. 4 it becomes clear that the $\mathcal{G}_L$ hides more information from p_{10} that what $\mathcal{G}_S$ does. This is due to the $\mathcal{G}_S$ limiting battery usage with a set of battery energy quotas for different S_i, including the choice to not protect a 59.92% of consumption ("Fridge" and "Others" in Table I).

Still, in becomes clear in Fig. 5 that the $\mathcal{G}_S$ can effectively protect an individual S_i. This is because: a) a smaller stored energy quota may still sufficiently absorb a significant percentage of consumption changes as long as the battery can satisfy the algorithm's demand for S_i allocated charge/discharge battery power and b) events in individual S_i are sparse allowing small battery quotas to refill. This is highlighted in the clustering results where the method ignores over 99% of values (that is 37 065 over 37 440 values).

A further observation is made by comparing the cluster similarity with the regression metric result for dp_7 in Table IV. This is (33.33%, 14.29%) for the former and (94.49%, 99.99%) for the latter, for (B1, B2) respectively. The low index for cluster similarity signifies the tendency of $\mathcal{G}_S$ to map the majority of $p_7(t)$ values to $p'_7(t)$ values within the same cluster. This is due to the sparsity of occurring consumption events. This is shown in Fig. 5 which compares a segment of dp_1 (laptop operation) and dp_7 (washer) with their modified traces. However, the regression protection index is high. This is due to getting a very large number of nonzero $dp_7(t)$ values, a very large number of

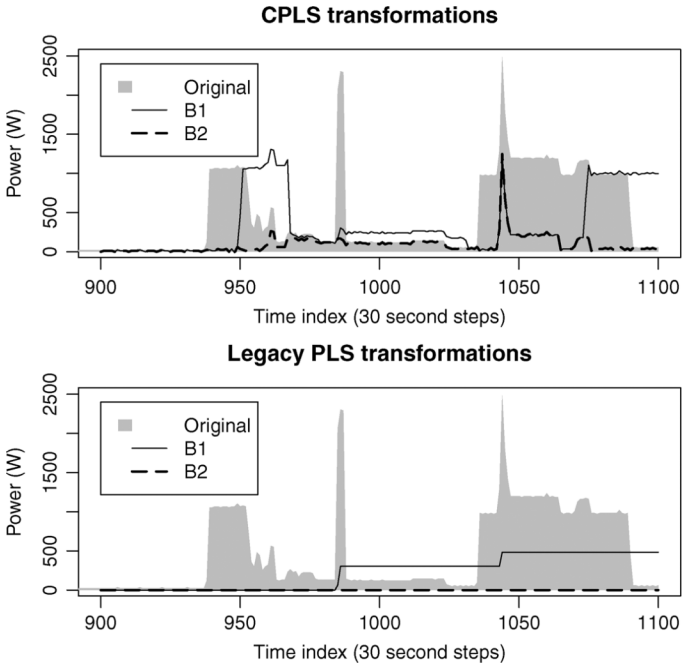


Fig. 4. Example of impact of selective PLS transformations to CPLS and comparison with legacy transformation in [9].

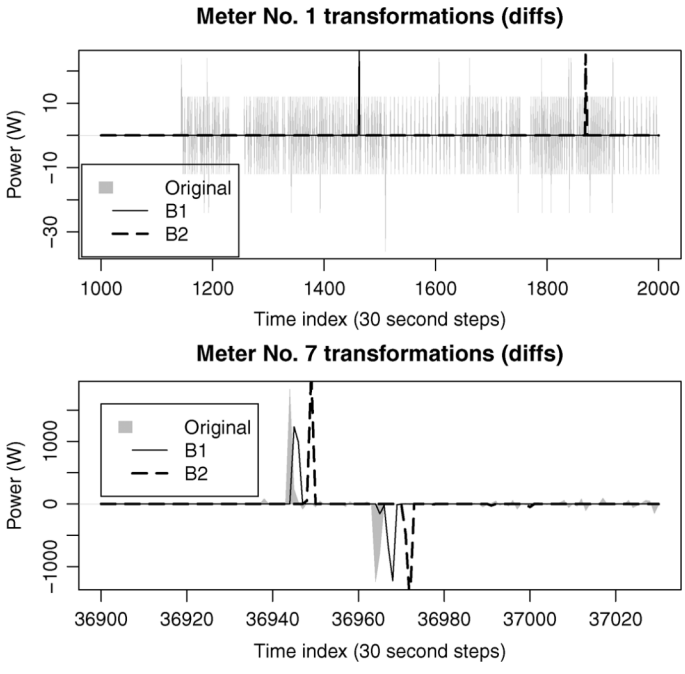


Fig. 5. Examples of detected edges following selective PLS privacy transformations on meters No. 1 and No. 7.

zero $dp'_7(t)$ values, and $\bar{dp}_7 \approx 0$, yielding $SS_E \gg SS_R$ and thus $R^2 \to 0$. That is, in this case, clustering highlights the algorithm's failure to protect the occurrence of a class of events demanding high power consumption, whereas regression highlights the algorithm's ability to protect the finer consumption details of a particular appliance.

On a more general note, we consider that the three given privacy metrics evaluate different aspects of privacy: the information theoretic and cluster analysis metrics capture the probabilistic characteristics of a certain dataset, whereas regression captures the manner with which values change in time. For example, consider the case where events are shifted in the time domain; this will change (permute) the information in time but not its probabilistic space. Hence, the variational distance and cluster similarity metrics become more suitable when we prefer to protect the information associated with the occurrence of an event (or a cluster of similar events), rather than the time of its occurrence; whereas regression becomes more suitable when the privacy of the "timing information" is also important. To this end, the existence of multiple metrics allow us to "attack" the algorithm in manifold ways.

E. ElecPrivacy Costs and Impact

An ElecPrivacy system and algorithm, such as $\mathcal{G}_{\mathcal{S}}$ and $\mathcal{G}_{\mathcal{L}}$ discussed above, can mask the operation of appliances (and protect home privacy) by modifying home usage consumption patterns. This may induce a cost to the customer and impact the utility, depending on the battery economics, energy generation costs, and pricing. In this section, we briefly compare a number of system performance factors for $\mathcal{G}_{\mathcal{S}}$ and $\mathcal{G}_{\mathcal{L}}$, based on the work in [23]. We consider the following performance factors.

- Battery *cycle life efficiency* (CLE): a certain charge/discharge battery pattern modifies the expected (estimated) battery cycle life. We suggest that the cycle life can be estimated as the average $\bar{L} = (1)/(c)\sum_{i=1}^{c} L(d_i)$, where $L(d_i)$ is the estimated battery cycle life for a given cycle pattern, d_i is the battery *depth of discharge* (DOD) of the ith cycle, and c is the total number of occurred cycles. We can calculate the CLE $\mathcal{F}$ as $\mathcal{F} = \bar{L}/L_{\max}$, where $L_{\max}$ is the maximum battery life. We calculate $L_{\max}$ using a logarirthic model from [24].
- *Battery energy throughput* (BET): we define BET as the total energy a battery has provided through discharging, and we consider that BET provides an index of the battery energy losses due to charge/discharge, including thermal transference.
- Peak average power ratio (PAPR): we define the PAPR as $\mathcal{C} = p_{\max}/p_{\rm rms}$, where $p_{\rm rms}$ is the *root mean square* of all $p(t)$ values. We consider that PAPR can be used to study the impact of ElecPrivacy to the utility as lower PAPR corresponds to more efficient energy utilization and reduced power peaks. (This is based on the fact that p_i, $\mathcal{G}_L(p_i)$ and $\mathcal{G}_S(p_i)$ have approximately the same average.)

The results for the above metrics are given in Table V. We observe that the legacy algorithm $\mathcal{G}_L$ has smaller CLE, smaller PAPR, and larger BET as compared to the selective algorithm $\mathcal{G}_S$. That is, $\mathcal{G}_S$ uses the battery more efficiently and it uses less battery energy; on the downside, $\mathcal{G}_S$ does not improve PAPR significantly, considering that, in this case, the PAPR without the use of any battery is 14.7. We also observe that the larger battery (B2) will increase BET and reduce PAPR. That is, a larger battery may help towards load shedding on the expense of increased battery energy losses.

Generally, selectiveness helps the battery to better maintain total energy charge levels even in cases where the algorithm forces the battery to use large (dis-)charge power. This is shown in Fig. 6 which helps us to compare the battery charge level fluctuations due to privacy transformations between $\mathcal{G}_{\mathcal{S}}$ and $\mathcal{G}_{\mathcal{L}}$.

TABLE V
CLE, BET, AND PAPR RESULTS FOR TRANSFORMATIONS $\mathcal{G}_{\mathcal{L}}$ AND $\mathcal{G}_{\mathcal{S}}$ USING BATTERIES B1 AND B2

Metric	Legacy $\mathcal{G}_{\mathcal{L}}$		Selective $\mathcal{G}_{\mathcal{S}}$	
	B1	**B2**	**B1**	**B2**
CLE	0.96	0.97	0.99	0.99
BET (Wh/day)	1400	1839	490	547
PAPR	10.4	10.2	14.6	14.1

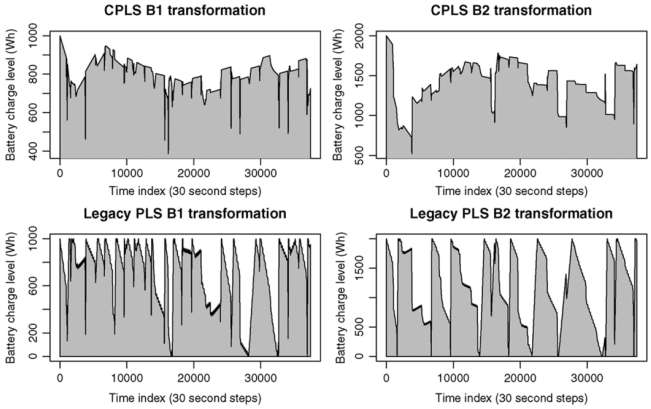


Fig. 6. Battery charge level fluctuations of selective privacy transformations and the legacy transformations in [9].

The legacy algorithm forces the (same) batteries into deeper charge-discharge cycles which limits the potential of battery energy storage or usage for alternative purposes.

A coprehensive analysis of the costs and the benefits of ElecPrivacy for the user, and the impact of ElecPrivacy on the grid is a nontrivial problem. For example, if ElecPrivacy is employed by a large number of houses, then the dynamics of the electrical network will be influenced. The level of influence depends on the size of the energy generation economics, the size of the network, and the specifics of the operating ElecPrivacy systems. An initial study of this problem can be found in [25]; we leave further analysis for future work.

Finally, we note that there is an analogy between the use of stored energy in electrical networks and the use of buffers in data network. The ability of the battery to maintain a buffer of energy for use in ElecPrivacy is reflected in the total energy consumed during the 13 days, which is 38092 Wh, 38045 Wh, 38059 Wh, 37818 Wh, and 37736 Wh; for CPLS, $\mathcal{G}_L$ with B1, $\mathcal{G}_L$ with B2, $\mathcal{G}_S$ with B1, and $\mathcal{G}_S$ with B2, respectively.

V. CONCLUSION

Home energy management algorithms can help reduce the exposure of sensitive energy usage information. Analysis of extensive metering data gathered from an apartment has shown that: a) expected consumption events can be predicted quite well and b) rechargeable battery resources can help protect privacy of particular appliances. Further, we have extended prior privacy metrics and we have shown that battery resources are more effectively used when detected power events are protected in a selective manner, as compared to an algorithm introduced in [9]. Thus, customized user control is key to enabling more secure and efficient battery usage. For example, the user may wish to hide the use of kettles, but not the use of a washing machine (whose prolonged energy usage may still allow NALM-like algorithms to identify it).

In future work the proposed privacy metrics will be further developed and "smarter" battery privacy algorithms will be designed. For example, appliance operation (TV, lighting, etc.) can be emulated in order to hide the occurrence of unusually prolonged periods of inactivity. Also, the physical implications of the system and its interaction with other energy objectives such as minimization of energy consumption and cost need to be studied further.

ElecPrivacy is not a substitute for smart metering data collection and usage regulations. In this respect, a smart metering system can be compared with other sensitive data collection systems such as healthcare systems. Different measures should coexist to offer a range of data privacy services. This may serve to increase the effectiveness of each solution.

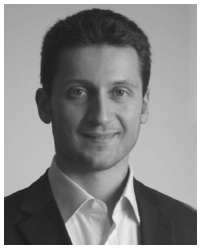
ACKNOWLEDGMENT

The authors would like to thank the Directors of the Toshiba Telecommunications Research Laboratory for their support.

REFERENCES

[1] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 52–62, 2009.
[2] C. Laughman, D. Lee, R. Cox, S. Shaw, S. B. Leeb, L. K. Norford, and P. Armstrong, "Advanced nonintrusive monitoring of electric loads," *IEEE Power Energy Mag.*, pp. 56–63, 2003.
[3] H. Y. Lam, G. S. K. Fung, and W. K. Lee, "A novel method to construct taxonomy electrical appliances based on load signatures," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 653–660, 2007.
[4] E. L. Quinn, "Privacy and the new energy infrastructure," Feb. 2009, available: [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1370731
[5] A. Lee and T. Brewer, "Smart grid cyber security strategy and requirements," NISTIR 7628, 2nd Draft, Feb. 2010.
[6] NIST, "Guidelines for smart grid cyber security: Vol. 2, Privacy and the smart grid," NISTIR 7628, Aug. 2010.
[7] Task Force Smart Grids, "Expert Group 2: Regulatory recommendations for data safety, data handling and data protection," Feb. 2011 [Online]. Available: http://ec.europa.eu/energy/gas_electricity/smartgrids/doc/expert_group2.pdf
[8] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Gaithersburg, MD, Oct. 4–6, 2010, pp. 238–243.
[9] G. Kalogridis, C. Efthymiou, T. Lewis, S. Denic, and R. Cepeda, "Privacy for smart meters: Towards undetectable appliance load signatures," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Gaithersburg, MD, Oct. 4–6, 2010, pp. 232–237.
[10] P. Vytelingum, T. D. Voice, S. D. Ramchurn, A. Rogers, and N. R. Jennings, "Agent-based micro-storage management for the smart grid," in *Proc. Auton. Agents MultiAgent Syst. (AAMAS)*, Toronto, ON, Canada, May 2010, pp. 39–46.
[11] Nedap, "The power router," Mar. 2011 [Online]. Available: http://www.powerrouter.com
[12] Toshiba, "Solar power generation," Mar. 2011 [Online]. Available: http://www.toshiba.co.jp/env/en/energy/solar.htm
[13] Plogg, "Plogg hardware," Mar. 2011 [Online]. Available: http://www.plogg.co.uk/ploggproducts.html
[14] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," 2010 [Online]. Available: http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.33.pdf
[15] D. Chaum, "The dining cryptographers problem: Unconditional sender and recipient untraceability," *J. Cryptol.*, vol. 1, no. 1, pp. 65–75, 1988.
[16] Y. Deng, J. Pang, and W. P. , "Measuring anonymity with relative entropy," in *Proc. 4th Int. Workshop Formal Aspects Security Trust*, Hamilton, ON, Canada, 2006, vol. 4691, Lecture Notes in Computer Science, pp. 65–79.

[17] A. Serjantov and G. Danezis, R. Dingledine and P. F. Syverson, Eds. , "Towards an information theoretic metric for anonymity," in *Proc. Workshop Privacy Enhancing Technol.*, 2002, vol. 2482, Lecture Notes in Computer Science, pp. 41–53.
[18] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. Berlin, Germany: Springer-Verlang, 1995.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 2006.
[20] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[21] R Development Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria, 2009, R Found. Stat. Comput. [Online]. Available: http://www.R-project.org
[22] Toshiba, "SCiB," Mar. 2011 [Online]. Available: http://www.scib.jp/en/product/spec.htm
[23] G. Kalogridis, F. Zhong, and S. Basutkar, "Affordable privacy for home smart meters," in *Proc. IEEE Int. Workshop Smart Grid Security Commun. (SGSC)*, Busan, Korea, May 26–28, 2011.
[24] L. H. Thaller and H. S. Lim, "A prediction model of the depth-of-discharge effect on the cycle life of a storge cell," Glenn Research Center, NASA, Tech. Rep. NAS3-22238, Jan. 1, 1987 [Online]. Available: http://hdl.handle.net/2060/19870012878
[25] S. Denic, G. Kalogridis, and F. Zhong, "Privacy vs. pricing for smart grids," in *Proc. 1st Int. Conf. Smart Grids, Green Commun. IT Energy-Aware Technologies, ENERGY'11*, Venice, Italy, May 22–27, 2011, pp. 153–158, IARIA.



Georgios Kalogridis received the Diploma degree in electrical and computer engineering from the University of Patras, Greece, in 2000 and the M.Sc. degree in advanced computing from the University of Bristol, U.K., in 2001. He is currently working toward the Ph.D. degree in mathematics from Royal Holloway, University of London, U.K.

He is a Senior Research Engineer at Toshiba Telecommunications Research Laboratory in Bristol, UK. His research interests include information security and privacy, network reliability, optimization, mobile agents, combinatorial mathematics, and smart grid communications. He is the inventor of numerous patents granted in the U.K. and worldwide, and he has been actively involved in collaborative projects and ETSI standardization activities.

Mr. Kalogridis is a member of IET and TCG.

Rafael Cepeda (S'96-M'01-SM'09) graduated as an electronics engineer from the Antonio Nariño University, Bogotá, Colombia, in 1998, and received the M.Sc. and Ph.D. degrees in electrical and electronic engineering from the University of Bristol, U.K., in 2001 and 2009, respectively.

He is currently a Principal Research Engineer at Toshiba Research Europe Limited's Telecommunications Research Laboratory (TRL), Bristol. His current research interests include the study of wireless channel propagation aspects and multiple antenna technology to optimize wideband communication systems.

Stojan Z. Denic (S'04–M'06) received the B.Sc. and M.Sc. degrees in electrical engineering from the Faculty of Electronic Engineering, University of Nis, Nis, Serbia, in 1994 and 2001, respectively, and the Ph.D. degree in electrical engineering from the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, in 2006.

From 2006 to 2008, he was with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, as a Research Assistant Professor. He currently holds a Senior Research Engineer post with the Telecommunications Research Laboratory, Toshiba Research Europe Limited, Bristol, U.K. His research interests include information theory, communication over uncertain channels, smart grid communications, control over communication channels, optical communications, and coding for wireless and digital recording channels.

Tim Lewis received the Ph.D. degree from the University of Edinburgh, U.K., in 2005.

He is a Principal Research Engineer at Toshiba Telecommunications Research Lab, Bristol, U.K., where he leads a team of researchers working in the area of wireless network protocol optimization. He has published in the research areas of parallel compilers, evolutionary optimization, protocol optimization and smart grid communications, and holds several patents in these fields.

Costas Efthymiou (S'99–M'99) received the M.Sc. degree in communications systems and signal processing and the M.Sc. (Research) degree in wireless communications systems from the University of Bristol, U.K., in 2001 and 2009, respectively.

He worked at the Telecommunications Research Laboratory of Toshiba Research Europe Ltd. between 2005 and 2010 as a Senior Research Engineer in Wireless Communications and has been involved in a number of EU and U.K. Technology Strategy Board projects such as GOLLUM, ARAGORN, and ViewNet. His research interests include link characterization, dynamic optimization techniques for wireless networks, smart metering networks, and security/privacy in smart grids.

Mr. Efthymiou is a member of the IET.