KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 4 (2018) 35–41

www.elsevier.com/locate/icte

Towards quantitative evaluation of privacy protection schemes for electricity usage data sharing

Daisuke Mashima[a,*], Aidana Serikova[b], Yao Cheng[c], Binbin Chen[a]

[a] *Advanced Digital Sciences Center, Singapore*
[b] *Nazarbayev University, Kazakhstan*
[c] *Institute for Infocomm Research, A*STAR, Singapore*

Received 1 December 2017; received in revised form 17 January 2018; accepted 18 January 2018
Available online 7 February 2018

Abstract

Thanks to the roll-out of smart meters, availability of fine-grained electricity usage data has rapidly grown. Such data has enabled utility companies to perform robust and efficient grid operations. However, at the same time, privacy concerns associated with sharing and disclosure of such data have been raised. In this paper, we first demonstrate the feasibility of estimating privacy-sensitive household attributes based solely on the energy usage data of residential customers. We then discuss a framework to measure privacy gain and evaluate the effectiveness of customer-centric privacy-protection schemes, namely redaction of data irrelevant to services and addition of bounded artificial noise.

Keywords: Privacy; Smart meter data; Quantitative evaluation

1. Background

Thanks to the penetration of smart meters and other types of commodity electricity usage monitoring devices, availability of fine-grained electricity usage data has increased remarkably. Besides utilization by utility companies, for instance demand forecasting and fault/anomaly detection, such data may be shared with third-party service providers either directly from customers (e.g., an energy usage monitoring device may upload data to the service provider's cloud for data analytics, etc.) or via utility companies (e.g., by means of Green Button Connect My Data [1]) to benefit from a variety of services, including energy-saving recommendations, social gaming, and services like demand response.

On the other hand, we are facing a number of new types of privacy risks that were not found in the age prior to the smart grid era. Privacy concerns associated with residential energy usage data have been outlined by National Institute of Standards and Technology (NIST) [2] and include leakage of personally-identifiable information and behavioral information. Moreover, unlike power utility companies that are strictly bound by regulations, other service providers may have the freedom to utilize the collected data for unclaimed purposes and/or share the collected data or analysis results with another party, e.g., advertising or marketing companies, without explicit consent from customers. Therefore, it is not feasible for electricity customers to retain control and awareness over usage of their data once the data are released. Nevertheless, most electricity customers share their data without enough understanding privacy exposure or ways to mitigate such risks [2].

To allow electricity customers to control privacy risks upon sharing electricity usage data with other parties, a framework called customer-centric energy usage management was proposed [3]. This framework can accommodate a variety of data pre-processing schemes applied by customers themselves for privacy protection [4,5] and is well aligned with policies regarding privacy and data ownership established by utility

* Corresponding author.
E-mail address: daisuke.m@adsc.com.sg (D. Mashima).
Peer review under responsibility of The Korean Institute of Communications Information Sciences.

https://doi.org/10.1016/j.icte.2018.01.006

companies in the US, e.g., [6], as well as European Union [7]. However, they did not show any quantitative evaluation of privacy gains, which can provide electricity customers with meaningful guidelines regarding how much pre-processing is needed to attain the expected level of privacy.

In this paper, we first design mechanisms to estimate privacy-sensitive household information based on household-level energy usage data to highlight potential privacy risks through experiments using real-world energy usage traces [8]. We further discuss a way to measure privacy gains of two privacy-protection mechanisms by means of redaction and artificial noise, which are introduced in the context of the aforementioned customer-centric electricity usage data management [3,4].

The rest of this paper is organized as follows. In Section 2, we discuss the literature on privacy pertinent to electricity usage data. In Section 3, to educate electricity customers, we demonstrate the feasibility of identifying privacy-sensitive household information with only electricity usage data. In Section 4, we discuss a framework to measure privacy gains and apply it to evaluate the effectiveness of two types of privacy-protection schemes that electricity customers can apply to mitigate privacy risks. We provide supplementary discussion in Section 5 and then conclude the paper in Section 6.

2. Related work

Kavousian et al. [9] analyzed the determinants of household electricity usage. The results indicated that household characteristics, appliance, electronics stock, and occupants indeed have a large influence on residential electricity usage patterns. An Irish case study [10] also examined the correlation between household/occupant characteristics and electricity usage using a multiple linear regression model. Their results demonstrate that, in addition to household characteristics, household composition and status of the head of household (e.g., age and social class) also have a strong correlation with electricity usage, which has provided a foundation for our investigation.

Beckel et al. [11] used an electricity usage dataset that was collected during a smart meter trial. Along with the electricity usage data, users' responses to a questionnaire before and after the trial are available and include various household characteristics. Based on these ground truth data, the authors demonstrated the feasibility of revealing characteristics from electricity usage data using various classifier models with an overall accuracy of around 70%. This feasibility is further supported by Aderson et al. [12], who demonstrated a concept of energy monitoring for a smart census. Recently, Cong et al. [13] conducted work on discovering missing user attribute labels using smart meter data. In this work, we investigate how much sensitive information can be inferred without any privacy protection, which is based on the feasibility revealed by these efforts. We further introduce extra features to enrich the feature space and apply other data analysis techniques for better accuracy. Moreover, we consider this accuracy as a baseline and evaluate the effectiveness of privacy-protection schemes.

Based on the assumption that the power utility companies fulfill their duty to protect users' electricity usage data as the data custodian, the focus of privacy protection is shifting to data sharing with third-party service providers. In this direction, researchers have proposed customer-centric energy usage management, a privacy protection scheme to enable meaningful data sharing with third parties while preserving users' privacy [3]. We should note that, customer-centric energy usage data management does not aim at privacy protection against utility companies, but against third-party service providers. Thus, it is complementary to, for example, battery-based privacy protection schemes like [14,15]. Moreover, the framework is orthogonal to privacy protection against attackers targeting smart metering infrastructures, e.g., those summarized in [16]. While [3] implemented privacy protection by means of redaction, there is another work that proposed to add artificial noise before data sharing to mitigate privacy risks [4]. However, to the best of our knowledge, there is no quantitative evaluation regarding how much privacy gain is attained from these protection schemes, which has motivated us to carry out such a study.

3. Estimating privacy-sensitive household attributes based on energy usage data

3.1. Residential energy usage dataset

To design and evaluate baseline schemes to estimate privacy-sensitive household attributes, and eventually, to evaluate the effectiveness of privacy-preservation schemes in the next section, we utilize a publicly-available electricity usage dataset collected in the UK, called the Household Electricity Survey (HES) dataset [8]. The primary reason we chose this dataset is that, in addition to electricity usage data with either 10-min or 2-min granularity, this dataset includes various details of each subject household obtained through the survey, which will be discussed later in this section.

Regarding electricity usage data, we used measurements collected at 2-min intervals in 220 households. HES data consist of appliance-level electricity usage data, so we aggregated energy consumption of all appliances in each household to approximate household-level traces. Furthermore, in order to make the data closer to realistic smart meter data, we down-sampled the 2-min interval household-level traces into 10-min intervals. Finally, because the period of data collection differs among households, we normalized the data by using the overall average for each season to remove seasonality.

Among the household details available in the HES dataset, in this study we focused on the following, which are considered to have marketing value and are therefore privacy sensitive: whether the household is occupied by a single person (*Single*), size of household occupancy (*Occupancy*), employment status of a household head (*Employment_Status*), whether a household has any children (*Children*), and the social grade of each household (*Social_Grade*). Class labels were determined based on the data, and their definitions are summarized in Table 1. Namely, Single and Children are defined as boolean (i.e., true or false), Occupancy is set to 1 if the size of occupancy (i.e., the number of residents) is higher than 2 while it is set to 0 otherwise, and Employment_Status is defined as binary regarding whether the