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*

Abstract

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

Thanks to the penetration of smart meters and other types of commodity electricity usage monitoring devices, availability of fine-grained electricity usage data has been dramatically increased. Besides utilization by utility companies, such as 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]) for benefiting from a variety of services, such as energy saving recommendation, social gaming, services like demand response, and so forth.

On the other hand, we are facing a number of new types of privacy risks, which were not found in the age before smart grid. Privacy risks and concerns associated with residential energy usage data have been outlined by NIST [2], which include leakage of personally-identifiable information, behavioral information, and so forth. Moreover, unlike power utility companies that are strictly bound by regulations, other service providers may have the freedom to utilize the collected data for other, unclaimed purposes and/or share the collected data or analysis results to yet another party, e.g., advertisement 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 exposures or the way of mitigating 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], which can accommodate a variety of data pre-processing schemes applied by customers themselves for privacy protection [4, 5]. The proposed framework is well aligned with policies regarding privacy and data ownership established by utility companies in the US, e.g., [6] as well as European Union [7]. However, they did not show any quantitative evaluation of the privacy gain, which can provide electricity customers with meaningful guideline 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 for measuring privacy gain of two privacy-protection mechanisms by means of redaction and artificial noise, which are introduced in the context of aforementioned customer-centric electricity usage data management [3, 4].

The rest of this paper is organized as follows. In Section 2, we discuss literature on privacy pertinent to electricity usage data. In Section 3, to educate electricity customers, we demonstrate the feasibility to identify privacy-sensitive household information only with electricity usage data. In Section 4, we discuss a framework for measuring privacy gain and apply it to evaluate the effectiveness of two types of privacy-protection measures 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] analysed the determinants on the household electricity usage, the results of which indicate household characteristics, appliance, and 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 the electricity usage using a multiple linear regression model. Their results demonstrate that, besides house characteristics, household composition and status of household head (e.g., age, social class) also have a strong correlation with the electricity usage, which has provided foundation for our investigation.

Beckel et al. [11] use an electricity usage dataset which is collected during a smart meter trial. Along with the electricity usage data, users' responses to a questionnaire before and after the trial are available which include various household characteristics. Based on these ground truth data, they demonstrate the feasibility of revealing characteristics from electricity usage data using various classifier models with an overall accuracy around 70%. This feasibility is further supported by Aderson et al. [12] who demonstrate a concept of energy monitoring for smart census. Recently, Cong et al. [13] also conducted work on discovering missing user attribute labels using the 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 as well as 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 would do their duty to protect the users' electricity usage data as the data custodian, the focus of privacy protection is shifting to data sharing with third-party service providers. Towards 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, it is also orthogonal to privacy protection against attackers targeting smart metering infrastructure, e.g., ones summarized in [16]. While [3] implements privacy protection by means of redaction, there is another work introducing 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 data collected in the UK, called Household Electricity Survey (HES) dataset [8]. The primary reason why we chose this dataset is that, besides electricity usage data with either 10-minute or 2-minute granularity, this dataset also 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 a 2-minute interval in 220 households. HES data consist of appliance-level electricity usage data, so we aggregated energy consumption of all appliances for 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-minute interval household-level traces into 10-minute interval. Finally, because the period of the data collection differs among households, we normalized the data by using the overall average for each season to remove seasonality.

Table 1: Class definitions for each attribute

Attribute	Class	Definition	# of Samples
Single	1	Single	62
	0	Not Single	158
Occupancy	1	> 2	84
	0	≤ 2	136
Employment_ Status	1	Full-time	123
	0	Otherwise	97
Children	1	With children	72
	0	Without children	148
Social_Grade	1	"A" or "B"	76
	0	Otherwise	144

Among the household details available in the HES dataset, in this study we focused on the followings, which are considered to have marketing values and therefore privacy sensitive: whether a household is occupied by a single person or not (*Single*), how big is household occupancy (*Occupancy*), what is the employment status of a household head (*Employment_Status*), whether a household has any children or not (*Children*), and social grade of each household (*Social_Grade*). Class labels are decided based on the data and 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 bigger than 2 while it is set to 0 otherwise, and Employment_Status is defined as binary regarding whether full-time employed or not. In the HES dataset, the social grade has 6 levels (A, B, C1, C2, D, and E), and we grouped A and B, which correspond to the high social grade, and formed the other group for the rest.

3.2. Designing Baseline Classifiers

This section discusses the design of baseline classifiers that are assumed to be used by a curious (or malicious) third-party energy-data analytics service provider that attempts to reveal privacy-sensitive data of each customer.

We initially defined totally 114 features derived from the aforementioned energy usage data. Based on our preliminary experiment, features calculated based on weekly (i.e., 1-week long) data showed better accuracy overall compared to ones computed based on monthly data, thus the results discussed on this paper is based on the features computed using 1-week data. For the experiment in this section, we used the first week of data of each household, resulting in totally 220 samples. Our initial list of features included basic ones such as average, variance, and quantiles of electricity usage of each household, as well as features proposed in [11, 13]. In addition, we included features derived from time series analysis (e.g., autocorrelation, ARIMA degrees, kurtosis, skewness, etc.) and Fast Fourier Transform (e.g., the most dominant frequency).

Then, we performed feature selection by Random Forest-Recursive Feature Elimination (RF-RFE) [17] for each household attributes to be estimated. This feature selection method provides importance score for each feature, and according to the score, we first selected 15 features out of the population for each classification, which are summarized in Figure 1. Using these features, by using WEKA [18], we applied multiple different classifiers that are popularly used, namely AdaBoost, kNN, SVM, Random Forest, Bagging, and BayesNet. Because including all of 15 features did not result in the best accuracy, we tweaked the number of features (i.e., selected a different number of features from the top) and conducted experiments for each classifier. As a result, we found that features of bold font in Figure 1 provided the best accuracy. Some of the results are shown in Figure 2.

In the figures, accuracy is computed based on the number of correctly-classified samples through 5-fold cross validation on WEKA. Note here that, WEKA's cross-validation implementation applies stratification of data (i.e., the ratio of samples of both classes are roughly the same in all groups). The best classifiers for the household attributes of our interest are summarized in Table 2. Note again that for the best classifiers, features shown with bold fonts in Figure 1 are used.

Table 2: Best-performed classifies and accuracy

Household Attribute	Classifier	Accuracy (%)
Single	AdaBoost	79.09
Occupancy	Random Forest	73.18
Employment_Status	Bayes Net	72.72
Children	SVM	75.45
Social_Grade	Random Forest	70.00

As can be seen from the table, privacy-sensitive household attributes can be estimated with over 70% accuracy by only using electricity usage data, and therefore sharing the fine-grained electricity usage data should be considered as serious privacy risks for electricity customers. Comparing our results with literature [11], even though direct comparison is not completely fair owing to the differences in dataset and definition of attributes, our classifiers attained noticeably better performance (over 10% increase) in estimating Social_Grade, while having similar accuracy in Single, Employment_Status, and Children. In the rest of this paper, we assume these classifiers are utilized by curious (or malicious) third-party service providers, and the accuracy achieved here (seen in Table 2) is used as the baseline for comparison when we evaluate the effectiveness of privacy-protection schemes.

4. Evaluating Effectiveness of Customer-centric Privacy-protection Schemes

In this section, we evaluate the effectiveness of privacy-protection schemes developed for customer-centric energy usage data management and sharing schemes [3]. In particular, as two data pre-processing techniques that a customer can apply before data sharing, we focus on the redaction of data [3] and the addition of artificial noise [4].

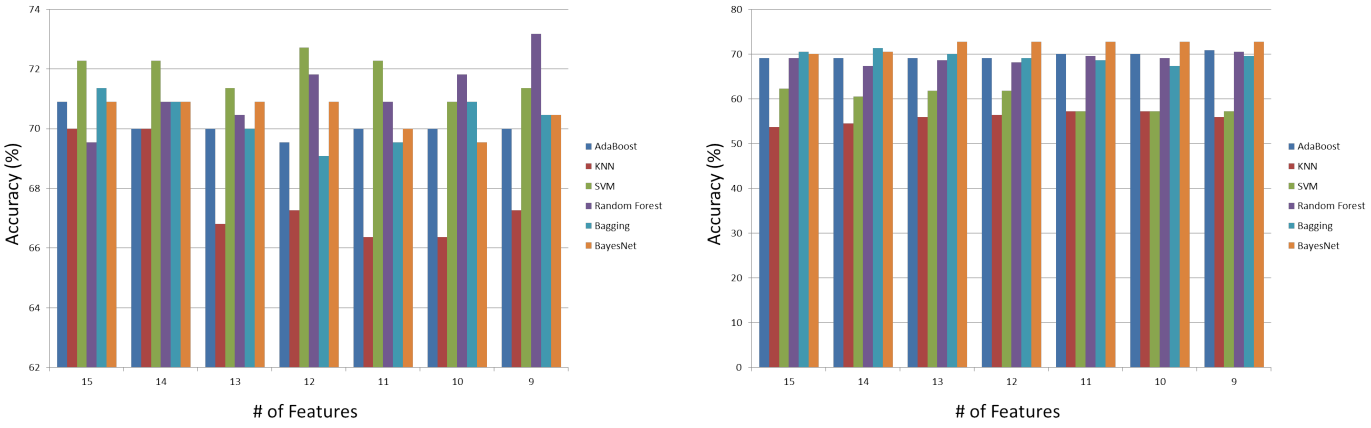
For the experiments in this section, we evaluate the effectiveness of privacy protection in the following way. For the sake of comparison with the baseline discussed in Section 3, we follow the similar procedure as 5-fold cross validation. Specifically, we randomly form 5 groups of samples in the stratified way just as done by WEKA in Section 3.2, and for each round, we use 4 of them for training and the other for testing. The difference from the typical 5-fold cross validation is that, while we use the original electricity usage data for training, for testing we use pre-processed data (see Figure 3). This way, we can compare the results with ones in Table 2. In sum, our experiments emulate a case where a (potentially malicious) service provider has classifiers trained based on original, labeled data collected from a number of customers and attempt to reveal privacy of customers who are submitting either original (Electricity Customer 1 in Figure 4 or pre-processed (Electricity Customer 2 in the same figure) electricity usage data respectively to evaluate the effectiveness of pre-processing for privacy protection.

4.1. Privacy Protection by Redaction

As discussed in [3], hiding some portion of data (e.g., showing only electricity usage during daytime) is considered effective for privacy protection. As can be seen in Figure 1, multiple classifiers rely on consumption during evening time as well as night time, which justify this approach. On the other hand, redacting part of data is considered still acceptable for many real-world services. For example, services like demand response services, which typically aim at controlling peak-time electricity demand and therefore are particularly interested in consumption during peak time in the afternoon [3].

Single	Occupancy	Employment_Status	Children	Social_Grade
1) **weekly_day_max**	1) **cons_we_day**	1) **weekly_eve_var**	1) **weekly_mor_var**	1) **ratio_weekly_eve_avg_noon_avg**
2) **weekly_total_max**	2) **weekly_eve_max**	2) **ratio_daily_mor_avg_noon_avg**	2) **weekly_day_var**	2) **ratio_var_we_avg_var_wd_avg**
3) **weekly_total_var**	3) **ratio_var_we_avg_var_wd_avg**	3) **ratio_cons_we_mor_cons_we_noon**	3) **ratio_cons_we_night_cons_we_day**	3) **ratio_daily_eve_avg_noon_avg**
4) **fft_dominant_freq**	4) **weekly_eve_var**	4) **fft_dominant_freq**	4) **weekly_eve_var**	4) **ratio_cons_we_noon_cons_wd_noon**
5) **weekly_day_var**	5) **ratio_cons_we_eve_cons_wd_eve**	5) **ratio_cons_we_noon_cons_we_day**	5) **ratio_cons_we_eve_cons_wd_eve**	5) **fft_dominant_freq**
6) **weekly_noon_percentile_25**	6) **ratio_cons_we_night_cons_we_day**	6) **ratio_cons_wd_mor_cons_wd_noon**	6) **ratio_var_we_avg_var_wd_avg**	6) **ratio_weekly_mor_avg_noon_avg**
7) **ratio_cons_we_eve_cons_wd_eve**	7) **weekly_day_var**	7) **ratio_weekly_mor_avg_noon_avg**	7) **ratio_daily_max_total_avg**	7) **ratio_cons_wd_night_cons_wd_day**
8) **weekly_morning_var**	8) **ratio_cons_we_eve_cons_we_noon**	8) **cons_wd_max**	8) **weekly_total_var**	8) **ratio_cons_wd_eve_cons_wd_noon**
9) **kurtosis**	9) **weekly_day_median**	9) **ratio_daily_eve_avg_noon_avg**	9) **ratio_cons_we_max_cons_we_min**	9) **ratio_weekly_noon_avg_total_avg**
10) **skewness**	10) weekly_total_var	10) ratio_cons_we_eve_cons_wd_eve	10) **fft_dominant_freq**	10) ratio_cons_wd_mor_cons_wd_noon
11) ratio_wd_avg_we_avg	11) cons_we_eve	11) ratio_cons_we_night_cons_we_day	11) **ratio_cons_we_noon_cons_wd_noon**	11) weekly_mor_median
12) weekly_noon_max	12) ratio_daily_night_avg_day_avg	12) ratio_var_we_avg_var_wd_avg	12) **var_of_daily_var**	12) ratio_daily_mor_avg_noon_avg
13) ratio_cons_we_eve_cons_we_noon	13) ratio_daily_mor_avg_noon_avg	13) ratio_cons_we_eve_cons_we_noon	13) **ratio_weekly_mor_avg_noon_avg**	13) ratio_daily_night_avg_day_avg
14) ratio_cons_we_noon_cons_wd_noon	14) fft_dominant_freq	14) var_of_daily_var	14) ratio_weekly_eve_avg_noon_avg	14) var_we_avg
15) weekly_day_percentile_75	15) ratio_cons_we_noon_cons_we_day	15) ratio_cons_wd_max_cons_wd_min	15) ratio_cons_wd_mor_cons_wd_noon	15) autocorr_daily

Figure 1: Short-listed features for each household attribute classification. The ones highlighted with bold font are the features used by the best classifiers.



(a) Occupancy (b) Employment_Status

Figure 2: Accuracy comparison among different classifiers with different numbers of features

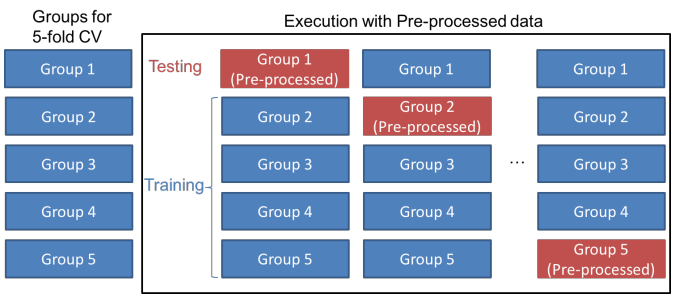


Figure 3: Evaluation of privacy gain following 5-fold cross validation using pre-processed data

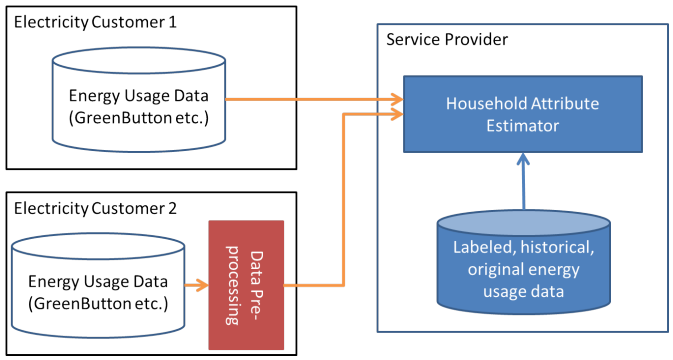


Figure 4: Our model for evaluating privacy gain. Our framework measures privacy gain in terms of difference in estimation accuracy against Electricity Customer 1, which shares original data, and one against Electricity Customer 2, which implements customer-centric data pre-processing before data sharing.

We performed two sets of experiments with different degree of redaction, one redacting electricity usage data except for typical peak hours (10am to 2pm) on each day and the other redacting data except 6am-6pm. We

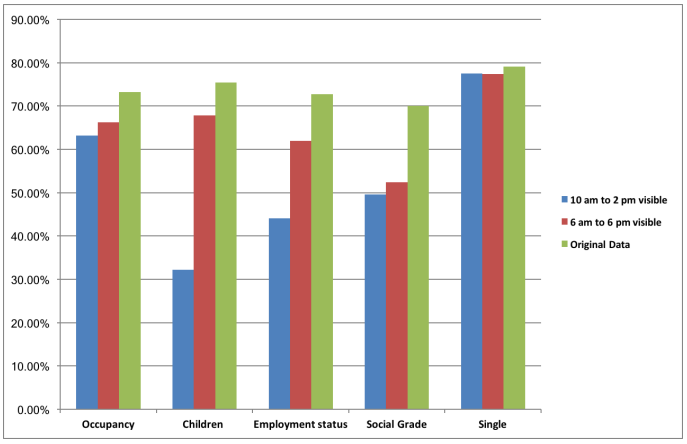


Figure 5: Accuracy of classification using redacted data

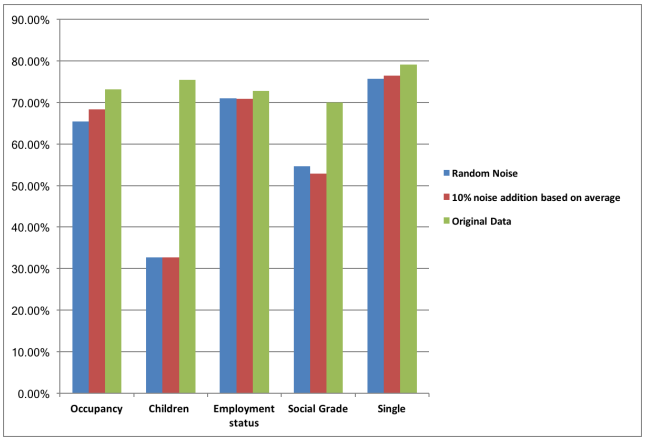


Figure 6: Accuracy of classification using data with artificial noise

assume that redacted data are replaced by the service provider with the overall average computed based on training data. The results are presented in Figure 5, along with the baseline accuracy from Table 2, which are labeled "Original Data". As can be seen, the accuracy decreases according to the degree of redaction. In particular, accuracy reduction (i.e., privacy gain) is significant in Children, Employment_Status, and Social_Grade.

4.2. Privacy Protection by Artificial Noise

Another privacy-protection strategy is to add an artificial, bounded noise to mask the exact electricity usage. Adding a noise would not be preferred for services that require exact data, such as electricity billing and performance evaluation of demand response services. However, a certain amount of noise is considered acceptable for energy-saving recommendation services etc. because approximate data are usually sufficient for many residential customers.

In this direction, we evaluated the effectiveness of bounded, randomly-added noise on electricity usage measurement in each time slot. Figure 6 shows the results of experiments with two different types of artificial noise. The first strategy is to add zero-mean, $\pm 10\%$ random noise (i.e., we generated random numbers between 0.9 and 1.1 for each electricity usage measurement and multiplied the factor with the corresponding measurement). The second strategy is slightly more intelligent and adds positive noise when the actual electricity usage of a certain time slot is below the overall average of the household while adding negative noise otherwise. As can be seen in the figure, we see noticeable decrease in classification accuracy for Children and Social_Grade.

However, compared to the redaction discussed in the previous section, overall the privacy gain by artificial noise seems limited. One of the plausible reason is that the added noise might have been to some extent canceled out when computing features based on the sum of measurements. If we consider further advanced mechanism to add a noise, the impact would be more noticeable. Moreover, the primary motivation for the artificial noise discussed in [4] was to make non-intrusive load monitoring (NILM) or load disaggregation [19, 20] techniques less accurate. In particular, NILM techniques often rely on "load signatures" derived from energy consumption patterns of each appliance, and noise on electricity usage data would make the signature matching less accurate. Therefore, when the feature set for classification includes those derived based on NILM results (e.g., usage pattern or frequency of a certain type of appliance), the privacy gain could be more significant.

5. Discussion and Future Research Directions

Based on the results shown in Figures 5 and 6, we can define a privacy-gain metric that summarizes the results for the sake of easier interpretation. For instance, we can calculate the (weighted) average of accuracy decrease. Alternatively, from customers' perspective, another metric could be defined in terms of how much information can be correctly identified. Exploration of effective metrics will be part of our future work.

In this study, we assumed that labeled dataset for training is given. It may be argued that this assumption would not be realistic because even utility companies don't know customer information other than basic information such as a name of household head, mailing address, phone number, and billing information. However, on the other hand, there are a non-negligible number of customers who may voluntarily surrender privacy-sensitive information, including ones we evaluated in this paper, along with their electricity usage data, through questionnaires requested in exchange for some benefits (e.g., discount or promotional coupons). After collecting data in such a way, a service provider would be able to collect a labeled dataset of sufficient size in reality.

One limitation of our study is that we did not take into account the adaptation of data analytics mechanism.

A service provider may adjust the feature set and/or classifiers to better handle pre-processed data (e.g., noisy data or redacted data). In other words, after somehow collecting a sufficient number of pre-processed data and ground-truth class labels, classifiers could be trained with them. Such a study is part of our future work.

Besides, we simplified the problem into a binary classification for all household attributes of interest. For example, regarding the occupancy size, instead of estimating the actual number, we in some sense just paid attention to identifying whether it is an extended family or not. In general, it is more challenging to estimate exact numbers, as also pointed out in [11]. Although we admit that it is an important part of our future work, the binary information explored in t his paper still has values for marketing and advertisement purposes.

Another direction for future work is to evaluate classifiers that include advanced features like ones derived from non-intrusive load monitoring etc. It is expected that households with different attributes would have different appliance usage patterns. Given the availability of open-source tool like NILMTK [20], derivation of such information would become feasible.

6. Conclusions

In this paper, we demonstrated the feasibility of estimating privacy-sensitive household attributes that could be potentially abused for unsolicited advertisement etc. Based on our experiments using a public dataset, all of 5 privacy-sensitive attributes considered in this paper can be estimated with over 70% accuracy. We further quantitatively studied the effectiveness of two privacy-protection measures that customers can practically apply before date sharing with potentially malicious third parties, namely redaction and artificial noise.

We hope our contributions shed lights on not only the privacy risks associated with electricity usage data but also the quantitative evaluation of privacy-protection schemes to counter such risks as well as to better educate electricity customers.

Acknowledgment

This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore under the Energy Programme and administrated by the Energy Market Authority (EP Award No. NRF2014EWT-EIRP002-040) and is also partly supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

[1] The green button, [Online]. Available: `http://www.greenbuttondata.org/`, (Date last accessed on Sep. 22, 2017).

[2] NIST Smart Grid, Guidelines for smart grid cyber security: Vol. 2, privacy and the smart grid, Guideline, Aug.

[3] G. Lahoti, D. Mashima, W.-P. Chen, Customer-centric energy usage data management and sharing in smart grid systems, in: Proceedings of the first ACM workshop on Smart energy grid security, ACM, 2013, pp. 53–64.

[4] D. Mashima, A. Roy, Privacy preserving disclosure of authenticated energy usage data, in: Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on, IEEE, 2014, pp. 866–871.

[5] D. Mashima, Authenticated down-sampling for privacy-preserving energy usage data sharing, in: Smart Grid Communications (SmartGridComm), 2015 IEEE International Conference on, IEEE, 2015, pp. 605–610.

[6] Privacy policy, [Online]. Available: `https://www.pge.com/en_US/about-pge/company-information/privacy-policy/privacy-policy/privacy-policy.page`, (Date last accessed on Oct. 9, 2017).

[7] Report on the local and regional consequences of the development of smart grids, [Online]. Available: `http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A7-2014-0019&language=EN`, (Date last accessed on Oct. 9, 2017) (2014).

[8] J. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, C. Evans, Household electricity survey: A study of domestic electrical product usage, Intertek Testing & Certification Ltd.

[9] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy 55 (2013) 184–194.

[10] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study, Energy and Buildings 48 (2012) 240–248.

[11] C. Beckel, L. Sadamori, T. Staake, S. Santini, Revealing household characteristics from smart meter data, Energy 78 (2014) 397–410.

[12] B. Anderson, S. Lin, A. Newing, A. Bahaj, P. James, Electricity consumption and household characteristics: Implications for census-taking in a smart metered future, Computers, Environment and Urban Systems 63 (2017) 58–67.

[13] Y. Cong, G. Sun, J. Liu, H. Yu, J. Luo, User attribute discovery with missing labels, Pattern Recognition.

[14] G. Kalogridis, C. Efthymiou, S. Z. Denic, T. A. Lewis, R. Cepeda, Privacy for smart meters: Towards undetectable appliance load signatures, in: Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on, IEEE, 2010, pp. 232–237.

[15] Z. Zhang, Z. Qin, L. Zhu, J. Weng, K. Ren, Cost-friendly differential privacy for smart meters: exploiting the dual roles of the noise, IEEE Transactions on Smart Grid 8 (2) (2017) 619–626.

[16] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang, A survey on privacy-preserving schemes for smart grid communications, arXiv preprint arXiv:1611.07722.

[17] P. M. Granitto, C. Furlanello, F. Biasioli, F. Gasperi, Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products, Chemometrics and Intelligent Laboratory Systems 83 (2) (2006) 83–90.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (1) (2009) 10–18.

[19] G. W. Hart, Nonintrusive appliance load monitoring, Proceedings of the IEEE 80 (12) (1992) 1870–1891.

[20] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava, Nilmtk: an open source toolkit for non-intrusive load monitoring, in: Proceedings of the 5th international conference on Future energy systems, ACM, 2014, pp. 265–276.