IoT-Privacy: To Be Private or Not To Be Private

Arijit Ukil
Innovation Lab
Tata Consultancy Services
Kolkata, India
arijit.ukil@tcs.com

Soma Bandyopadhyay
Innovation Lab
Tata Consultancy Services
Kolkata, India
soma.bandyopadhyay@tcs.com

Arpan Pal
Innovation Lab
Tata Consultancy Services
Kolkata, India
arpan.pal@tcs.com

***Abstract*—Privacy breaching attacks pose considerable challenges in the development and deployment of Internet of Things (IoT) applications. Though privacy preserving data mining (PPDM) minimizes sensitive data disclosure probability, sensitive content analysis, privacy measurement and user's privacy awareness issues are yet to be addressed. In this paper, we propose a privacy management scheme that enables the user to estimate the risk of sharing private data like smart meter data. Our focus is to develop robust sensitivity detection, analysis and privacy content quantification scheme from statistical disclosure control aspect and information theoretic model. We depict performance results using real sensor data.**

Keywords—privacy; smart meter; statistical disclosure; sensitivity; Wasserstein distance;

I. Introduction

IoT applications like smart home, smart energy management render innumerable benefits to human society. Sensors like smart meters collect sensitive personal information like detailed household energy consumption profile. However when such data is released to third parties, possibility of unintentional or malicious privacy breach like activity detection of the users is very high. Though released data can be privacy protected by standard privacy preservation techniques like PPDM using noise addition, suppression; sensor data owner should also be aware of the privacy content. Our analysis also acts as a precursor in enforcing optimal privacy preservation. In this paper, we present a privacy management scheme that detects, analyzes sensitive content of time-series sensor data and measures the amount of privacy. Without loss of generality, we consider smart meter data for investigation. We assume appliance load change recovery and peak load monitoring as privacy threats and main privacy violators. Intuitively, the anomalous events that are unpredictable and have the potential to arouse curiosity or undue interest are considered sensitive or private. Our scheme has two distinct components. First, we detect and analyze the sensitivity in typical sensor dataset using robust statistical method. Then we measure the privacy content of the sensor dataset using information theoretic model that is consistent with analytical reasoning. We perform experimental tests using real sensor data [1] and compare our results with relevant techniques [2-3].

II. Senstivity Detection and Analysis

We define sensitivity as statistical anomalies in the sensor data that describe the presence of unusual or unanticipated events. Most of the relevant proposed schemes are either completely dependent on the application use case (intrusive) [4] or totally independent [2 – 3] (rudimentary). We propose an unobtrusive sensitivity detection and analysis algorithm that discovers the anomaly points in sensor dataset ($\mathcal{S}_t$) while optimizing the masking and swamping effects. Firstly, the distribution pattern of $\mathcal{S}_t$ is derived using fourth order statistical moment (kurtosis, κ). When $\kappa > 3$ (leptokurtic), Rosner filtering is executed to minimize swamping effect [5 – 6]. Unlike in traditional outlier detection tests or clustering algorithms, we need not specify the number of sensitive points apriori. Given the upper bound, Φ, Φ number of backward selection tests are performed to identify initial anomalies, ν_i , i= 1, 2, …, Φ with respect to median deviation, suspected ν_i that maximize the deviation from underlying presumed student's-*t* distribution are considered as the sensitive points. When $\kappa \leq 3$, Hampel filter is employed to detect sensitive points so that masking effect is minimized. Hampel filter is a nonlinear data cleaning filter that identifies local outliers through Median Absolute Deviation scale estimator. We test sensitivity detection and analysis outcome using real smart meter data, REDD dataset [1] and the result is shown in fig. 1.

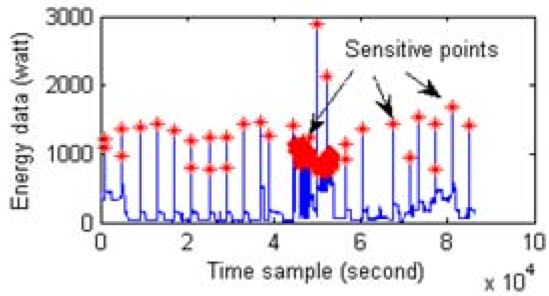

Fig. 1. Sensitivity detection and analysis

We prove that our proposed scheme has better detection capability with optimal masking and swamping effects when compared with [2-3] using different measures like Mahalanobis (ML) distance, Kullback Leibler (KL) divergence. For example, KL divergence of our scheme is more than that of [2– 3] and Sanov's theorem from large deviation theory shows more KL divergence means more detection capability. ML distance measure of the proposed method is higher than the relevant schemes [2 -3], depicting more detection ability.

III. Privacy Measurement and Quantification

Privacy measurement and quantification scheme is computed based on statistical and information theoretic model.

Our objective is to derive privacy measure from fundamental principle for disambiguation of privacy measure among different privacy preservation guarantees like *k*-anonymity, *l*-diversity. Below, we describe our proposed scheme. Functional block diagram is shown in figure 2.

A. Privacy measurement

Consider ν be the sensitive part of $\mathcal{S}$ (sensor data set) and λ be the non-sensitive part; $\mathcal{S} = \nu \cup \lambda$. We define privacy measure as the amount of difficulty to infer ν, when only λ is presented or how much the probability of finding ν, given $\mathcal{S}$, i.e. the information leakage transfer function $\Upsilon_{\mathcal{S},\nu}$: $\mathcal{S} \rightarrow \nu$ and $\Upsilon_{\mathcal{S},\nu} = \rho_M$ (privacy measurement value) $= \frac{\sum_{i=1}^{|\nu|} Pr(\nu_i) \log_2 \frac{1}{Pr(\nu_i)}}{\sum_{i=1}^{|\mathcal{S}|} Pr(\mathcal{S}_i) \log_2 \frac{1}{Pr(\mathcal{S}_i)}}$.

In [7], such metric is derived using mutual information $I(\mathcal{S},\nu)$. As $\nu \in \mathcal{S}$, leakage function $\mathfrak{L}_{\mathcal{S},\nu} = I(\mathcal{S},\nu)$ would indicate the maxima.

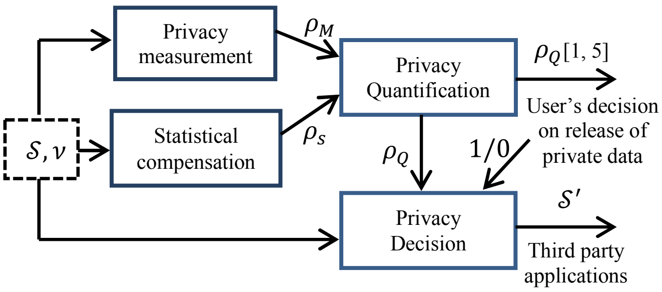

Fig. 2. Privacy measurement, quantification, user decision

B. Statistical compensation

In order to enhance privacy quantification accuracy, statistical compensation due to statistical relation between S and ν is computed. Here, two-sample Kolmogorov-Smirnov (KS) test of S and ν is performed. KS test is a non-parametric hypothesis test that evaluates the difference between the Cumulative Distribution Functions (CDFs) . It computes under the null hypothesis that S and ν are drawn from the similar distribution. When KS-test accepts null hypotheses, statistical compensation $\rho_S = 1$. When KS-test rejects null hypotheses, we propose L1-Wasserstein metric $(\mathcal{w}_{\mathcal{S},\nu})$ between $\mathcal{S},\nu$ to estimate statistical misfit or compensation as $\rho_S = \mathcal{w}_{\mathcal{S},\nu}$, that effectively deduces distribution dissimilarity in non-linear dynamic systems. Wasserstein distance quantifies the numerical cost with respect to distribution dissimilarity between pair of distributions, defined for $\mathcal{S},\nu$:

$$\mathcal{w}_{\mathcal{S},\nu} := \inf_{\mu \in \Omega(\mathcal{S},\nu)} \int_{\Omega} |x - y| \, d\mu(x,y),\ x \in \mathcal{S},\ y \in \nu$$

C. Privacy quantification

Logically, privacy quantification $\rho_Q = \rho_M \wedge \rho_S$. Algebraically, $\rho_Q = \rho_M \times \rho_S$. With $\rho_Q[0,1]$, we scale ρ_Q as $\rho_Q \mapsto \lceil \rho_Q \times 5 \rceil$: $\rho_Q = [1,5]$, with high magnitude of ρ_Q signifies more privacy risk probability in $\mathcal{S}$. Basically, for ease of understanding we provide privacy content quantification as integer value ranging between 1 and 5. We depict in figure 3 the outcome of privacy quantification of our method comparing with [2 -3]. We compute privacy quantification (privacy score) of four equal parts of the day. Comparing the privacy measured values in fig. 3 with sensitivity outcome in fig. 1, we observe that denser sensitive zone is assigned more privacy quantified value. Efficacy of the proposed scheme can be further established when an attack with standard disaggregation or NILM (Non-Intrusive Load Monitoring) is launched, which is our future research scope.

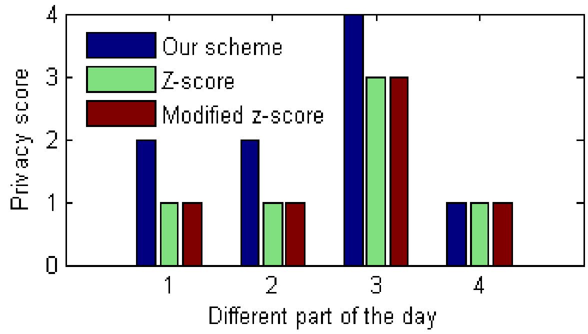

Fig. 3. Privacy measure outcome

D. Privacy decision

Privacy quantification enables the users to decide whether to share its private sensor data to avail third party applications. When affirmative, $\mathcal{S}'$ is computed and shared to third parties. Here, $\mathcal{S} \xrightarrow{\rho_Q} \mathcal{S}'$ using standard PPDM, where $\mathcal{S}'$ is the privacy preserved sensor data. Informally, the strength of perturbation or generalization of PPDM scheme is proportional to ρ_Q. For example, if $\mathcal{S}' = \mathcal{S} + \mathcal{N}$, $\mathcal{N} = f(\rho_Q)$ or *l*-diversity of $\mathcal{S}$ would be $l = g(\rho_Q)$.

IV. CONCLUSION

In this paper, we proposed a scheme in IoT systems that enables sensor data owner to assess the privacy risk of sharing its sensor data like smart meter data to third party applications in quantifiable terms. The proposed scheme is generic in nature and can adapt to different time-series sensor data based applications.

REFERENCE

[1] Z. Kolter, and M, J. Johnson, "REDD: A public data set for energy disaggregation research," SustKDD, 2011.

[2] R. Rao, S. Akella, G. Guley, "Power Line Carrier (PLC) Signal Analysis of Smart Meters for Outlier Detection," IEEE SmartGridComm, pp. 291 - 296, 2011.

[3] R. M. Nascimento, et al., "Outliers' Detection and Filling Algorithms for Smart Metering Centers ," IEEE PES , pp.1 - 6, 2012.

[4] W. Yang, et al., "Minimizing Private Data Disclosures in the Smart Grid," ACM CCS, pp. 412- 427, 2012.

[5] B.Rosner, "Percentage points for a generalized ESD many-outlier procedure," Technometrics, vol. 25, issue. 2, pp. 165 - 172, 1983.

[6] R. Serfling, and S. Wang, "General Foundations for Studying Masking and Swamping Robustness of Outlier Identifiers," Elsevier Statistical Methodology, August 2013.

[7] L. Sankar, S.R. Rajagopalan, S. Mohajer, and H.V.Poor, "Smart Meter Privacy: A Theoretical Framework," IEEE Transactions on Smart Grid, vol. 4, issue. 2, pp. 837 - 846, 2013.

[8] A. Halder, and R.Bhattacharya, "Further results on probabilistic model validation in Wasserstein metric," IEEE Annual Conference on Decision and Control, pp. 5542 - 5547, 2012.