# Personalized Differential Privacy Preserving Data Aggregation for Smart Homes

Xin-Yuan Zhang, Liu-Sheng Huang, Shao-Wei Wang, Zhen-Yu Zhu, Hong-Li Xu
School of Computer Science and Technology, USTC, Hefei, 230027, China
E-mail: dwz@mail.ustc.edu.cn, lshhuang@ustc.edu.cn, wangsw@mail.ustc.edu.cn, zzy7758@mail.ustc.edu.cn, xuhongli@ustc.edu.cn

***Abstract*-The aggregation of residents' private data drives improvements in the smart homes, however it comes with compromising on privacy. Hence, privacy preservation has become an increasing requirement for residents. Since users might have different privacy requirements, and their privacy requirements might be sensitive information, smart homes need a privacy preservation scheme to meet their demands. In this scheme, a user preserves the privacy of his/her data and privacy level locally by specifying his own privacy level in confidence, without trusting anyone else in smart homes. In addition, a user replies a single data element to the collector each time, instead of the whole dataset. It makes the scheme more challenging than the traditional centralized situation. In this paper, we propose a novel personalized local differential privacy preservation scheme for smart homes, which retains desirable utility while providing rigorous privacy guarantee.**

***Keywords-smart homes; aggregation; differential privacy; local; personalized***

## I. INTRODUCTION

With the rapid development of smart homes, it becomes inevitable to optimize resources and provide better service through aggregating data from users. For instance, smart homes can analyze the preferential temperature distribution utilizing the collected data, further ascertain the most commonplace temperature; in order to deliver appropriate services, it needs to discover the age structure of the residents. Nevertheless, dwellers' privacy requirement has been a vast obstacle to the widespread participation of contributing data to the collector [5]. Users would not like to share their sensitive information unless their privacy issues have been resolved appropriately. On the other hand, the collector will obtain a more accurate statistics estimation accompanied by the increasing contributors in smart homes. Due to the condition that the information from a single resident is not indispensable to analysis results, we can start from this point to preserve users' privacy while not comprising data utility too much. That is to say, we can find a scheme which fulfils the collection and analysis of sensor data, avoiding the privacy risks of participants' information simultaneously.

Differential privacy [12] is the state of the art privacy definition which aims to resolve the privacy risk in the process of data statistics. It has been certified that differential privacy can provide rigorous privacy guarantees, regardless of the background knowledge and computing power an adversary has. However, privacy preservation for smart homes belongs to a new differential privacy mechanism which is named local privacy [16]. The local privacy refers that the privacy protection algorithm is performed locally on residents' sensor devices. Another feature is that this mode doesn't need a trustable third-party. Moreover, the solicited privacy data is a piece of data, not a database. Therefore, the two common mechanisms of traditional differential privacy, that is the Laplace mechanism and the exponential mechanism [1], are not appropriate for the local mode. Since smart homes consist of multiple users with different privacy exceptions, a collector is confronted with a predicament which limits aggregation scheme. One possible solution is to set the global privacy budget to satisfy every user. This is likely to introduce too much noise that data utility will be extensively influenced. Another is to set a higher identical privacy budget to meet most of the users' demand. However, this option will also do harm to data utility. Therefore, we need to devise a personalized local differential privacy scheme which takes different privacy requirements of each resident into account. The personality means each contributed resident can specify themselves privacy budget, and retain the budget secretly. The aggregation will benefit a lot from the flexibility of personalized scheme.

In this paper, we propose a personalized privacy-preserving data analytics scheme for smart homes. In this scheme, we can maintain data utility and preserve privacy simultaneously. Specifically, this paper makes the following contributions:

- We formulate the privacy preservation issue in smart homes. In response to users' natural personalized privacy requirements, we design a novel personalized differential privacy scheme. It uses the binary randomized response mechanism as the building block. The scheme provides a better trade-off between privacy and utility. Apart from the scenario in smart homes, our scheme is also appropriate for other aggregation tasks.
- We propose a two-stage randomized response mechanism that satisfies the demand of real-time data collection. It also keeps residents' privacy budgets from any other nodes in smart homes, including the collector. In order to minimize the maximum estimation error, we devise a solution tailored to this problem. Moreover, our mechanism is lightweight to be deployed on local sensors.
- We conduct simulations to evaluate the performance of our proposed scheme. The simulation results not only show that the scheme provides a better trade-off between privacy and utility, but also reveal that the accuracy of estimation improves with the number and privacy budget of participates increasing.

The rest of this paper is structured as follows. Section II reviews the related work. Section III introduces the definition of the local differential privacy and the personalized local differential privacy. Section IV presents our randomized response algorithm. Section V discusses the privacy preservation framework in detail. Section VI shows our experimental results. In the end, Section VII concludes our work.

## II. RELATED WORK

### A *Privacy Data Aggregation for Smart Homes*

The aim of privacy preservation is to protect users' privacy from others, while collecting and analyzing private information. Privacy concerns from smart homes have been proposed in plenty of literatures (e.g. in [13], [14]). The work [6] proposed a framework which is based upon k-anonymity and role access control. Residents specify privacy levels according to the role of the data collector, then accomplish privacy preservation with k-anonymity algorithm. The work [9] introduced a dynamic method for modifying privacy in smart homes, based on the context using data hiding techniques to decrease the invasiveness, and retaining the functionality at the same time. However, these previous works can’t provide theoretical privacy guarantees as differential privacy [12].

### B *Local Differential Privacy*

The concept of local differential privacy was first proposed by S. P. Kasiviswanathan et al. in [16]. The authors of [4] proposed a scheme RAPPOR that addresses the problem of longitudinal data collection with local differential privacy. The work [17] presented a new mechanism named k-ary Randomized Response to estimate discrete distribution under local privacy. Reference [15] introduced general approaches to get minimax bounds under LDP.

### C *Personalized Privacy*

Jorgensen, Yu and Cormode [18] proposed a personalized differential privacy mechanism with a trustable data analyst. The algorithm of privacy protection is not performed on local clients. Wang et al. [5] introduced a data aggregation scheme that provides personalized privacy preservation for participants. However, the scheme can’t satisfy the need for real-time and multiple times data collection, and it also introduces too much noise.

To our best knowledge, there hasn’t been an existing work that applies personalized local differential privacy (PLDP) mechanism to data collection for smart homes. We design a scheme utilizes PLDP to protect users’ sensitive information longitudinally.

## III. PRELIMINARIES

### A. *Local Differential Privacy*

Local differential privacy (LDP) [16] is a rigorous privacy notion in local setting, which provides a more tough privacy guarantee than the traditional differential privacy. It is because the mechanism needn’t any trustable third-party. This is, users protect their private data from anyone else by themselves. They won’t report their private data to the collector unless their private data have been properly sanitized locally. The collector makes an analysis on received perturbed data. Despite owning the perturbed information, any adversaries can’t infer the users' true data, independent of their background knowledge and computational power. Another feature of local differential privacy is that the collected information is usually a single data element (e.g.., the time of sleep), not a dataset. Formally, local differential privacy is given below.

Definition 1 (Local Differential Privacy [16]): A randomized $\Gamma$ algorithm satisfies $\epsilon$-local differential privacy, if for all pairs of values m and $m' \in D$, and for all M $\subseteq$Range($\Gamma$),

$$P[\Gamma(m) \in M] \le \exp(\epsilon)\, P[\Gamma(m') \in M],$$

where $\epsilon$ denotes privacy budget, and $D$ is the domain of privacy data.

As a result, LDP is performed with a single data instead of a dataset, it provides more rigorous privacy protection than traditional differential privacy. However, LDP introduces too much noise because the size of $D$ is commonly very large. Hence, it's often difficult to achieve a fine trade-off between privacy and utility. Besides, users can't choose privacy levels by themselves. Because of these two drawbacks, we introduce a more effective personalized local differential privacy to overcome these two shortcomings.

### B. *Personalized Local Differential Privacy*

Personalized local differential privacy (PLDP) is a novel notion based on LDP, it satisfies different privacy requirements of users. Personalization means that users can specify their privacy budgets without sharing with others. The flexibility of PLDP makes more users be willing to participate in data aggregation tasks, and improves data utility. In order to further improve performance, we propose the concept of security domain [*s*, *e*] in this paper. The security domain refers to the minimum acceptable domain of a user's private data. For instance, in smart homes, the collector collects residents’ age information. By setting his/her security domain to [20, 30], a resident whose true age is 23 doesn't mind the replied value in [20, 30], but the collector doesn't know whether the data collected is the resident's real age. Moreover, users can specify their security domain according to their own requirements. In this manner, we can decrease the noise introduced by PLDP mechanism.

Definition 2 (Personalized Local Differential Privacy): A randomized $\Gamma$ algorithm satisfies ($s$, $e$, $\epsilon$)- personalized local differential privacy, if for all pairs of values m and $m' \in D$, and for all $M \subseteq Range(\Gamma)$,

$$P[\Gamma(m) \in M] \le \exp(\epsilon)\, P[\Gamma(m') \in M],$$

where [*s*, *e*] denotes security domain.

## IV. THE RANDOMIZED RESPONSE MECHANISM

### A. *Two-Stage Binary Randomized Response*

Two-stage binary randomized response(TBRR) consists of two detached but compact randomized responses on Bloom Filter [2]. Randomized response technique(RRT) [3] is an effective method to collect sensitive information while protecting the confidentiality of information. For instance, a user's private information is *m*, he/she flips a coin secretly, replies with *m* if the coin comes up head, otherwise select a value from the range of *m* randomly. Others cannot distinguish the received data from the true data *m*. In this way, the user has plausible deniability of any his/her answers. PLDP uses the feature of RRT to protect privacy locally.

Unfortunately, the privacy budget provided by RRT is ln(|*D*|+1), it will become larger with |*D*| increasing, *D* denotes the range of *m*, and |*D*| refers to the size of *D*. Lately the authors of [4] proposed an algorithm that maps a value m to a Bloom filter before applying randomized response strategy.

A Bloom Filter is comprised of a long binary vector and a series of pseudorandom hash functions. It's used to judge whether an element belongs to the specified set. Applying Bloom Filter into RRT, we can obtain privacy budgets independent of the value of |*D*|. TBRR is performed on local sensors in smart homes as the following steps with parameters (*d*, *s*, *t*, *p*, *q*).

*1) Initialize Bloom filter*

Map the user's value m to a position in a Bloom filter *B* using the selected hash function. The size of *B* is *d*.

*2) The first stage of TBRR*

First of all, set the corresponding security domain as [*s*, *e*]. For every bit *i* in B, if $s \leq H^{-1}(i) \leq e$ in *B*, apply the following RRT mechanism to create two new binary arrays $B_1$ and $B_2$:

$$B_{1i}=\begin{cases}1, & \text{with probability } p/2\\ 0, & \text{with probability } p/2\\ B_i, & \text{with probability } 1\text{-}p\end{cases} \text{ and}$$

$$B_{2i}=\begin{cases}1, & \text{with probability } p/2\\ 0, & \text{with probability } p/2\\ B_i, & \text{with probability } 1\text{-}p\end{cases},$$

where *p* is the probability argument specified by users, and it decides the user's privacy level. Meanwhile, it also reflects the personalization of TBRR.

*3) The second stage of TBRR*

Initialize a binary array *B'* with size d, set all its bits to 0. Then, for each bit in $B_1$ and $B_2$, apply the following RRT mechanism to modify the array $\mathrm{B}'$.

$$\mathrm{B'_i}=\begin{cases}1, & \text{with probability } 1, & \text{if } B_{1i}+B_{2i}=2\\ 0, & \text{with probability } 1, & \text{if } B_{1i}+B_{2i}=0\\ 1, & \text{with probability } q, & \text{if } B_{1i}\oplus B_{2i}=1\end{cases},$$

where *q* is the probability parameter assigned by collector. TBRR satisfies the demands of real-time and multiple times collections.

*4) Report*

Respond the perturbed binary array *B'* to the data collector in smart homes.

## *B. Differential Privacy of TBRR*

The TBBR mechanism protects privacy through introducing uncertainty using two different RRT mechanisms. We have proved our scheme satisfies **Definition 2**.

THEOREM 1: The Two-stage binary randomized response mechanism satisfies (*s*, *e*, $\epsilon$)-personalized local differential privacy,

$$\epsilon = \ln \frac{\mathrm{a}\cdot(1\text{-b})}{\mathrm{b}\cdot(1\text{-a})}. \tag{1}$$

Where $a=(1\text{-}0.5p)^2+pq(1\text{-}0.5p)$, and $b=0.25p^2+pq(1\text{-}0.5p)$.

*PROFF.* Let *x* be the original Bloom filter, *y* be the $B_1$, *z* be the $B_2$ and r be the reported array *B'* generated by TBBR Since (1) and (2), corresponding probabilities are

$$P(y_i=1|x_i=1) = 1\text{-}0.5p \text{ and}$$
$$P(y_i=1|x_i=0) = 0.5p \text{ when } s\leq H^{-1}(i)\leq e.$$
$$P(z_i=1|x_i=1) = 1\text{-}0.5p \text{ and}$$
$$P(z_i=1|x_i=0) = 0.5p \text{ when } s\leq H^{-1}(i)\leq e.$$

Then

$$P(r_i=1|x_i=1) =(1\text{-}0.5p)^2+pq(1\text{-}0.5p) =a$$
$$P(r_i=1|x_i=0) =0.25p^2+pq(1\text{-}0.5p) =b$$
$$P(r_i=0|x_i=1) =0.25p^2+p(1\text{-}q)\,(1\text{-}0.5p) =1\text{-}a$$
$$P(r_i=0|x_i=0) =(1\text{-}0.5p)^2+p(1\text{-}q)\,(1\text{-}0.5p) =1\text{-}b.$$

Without loss of generality, let *x* be *x**= {1, 0, 0, ..., 0}, the security domain be [1, *d*], the conditional probability is

$$P(B'=r|B=x^*) =a^{r_1} \times (1\text{-}a)^{1\text{-}r_1} \times \mathrm{b}^{r_2} \times (1\text{-}b)^{1\text{-}r_2} \times \cdots\times b^{r_d}\times(1\text{-}b)^{1\text{-}r_d}.$$

Based on [4], the ratio of two conditional probabilities is

$$\frac{P(B' \in M^*|B=A_1)}{P(B'\in M^*|B=A_2)} \leq \max_{B'_i\in M^*} \frac{P(B'=B'_i|B=A_1)}{\mathrm{P}(B'=B'_i|B=A_1)} = \frac{a\cdot(1\text{-}b)}{b\cdot(1\text{-}a)}.$$

Consequently, the TBRR mechanism has been proved it can provide (*s*, *e*, $\epsilon$)- personalized local differential privacy.

In smart homes, the collector needs to aggregate data from sensors real-time. An adversary can obtain abundant reports from a resident, this might lead risks to the user's privacy. To defense underlying longitudinal attacks, we have to apply randomized response twice. In the first stage of TBRR, we use two Bloom filters to increase the uncertainty of users' responses. The TBRR mechanism makes it more difficult for attackers to infer users' private information. It can satisfy the demand for real-time data collection in smart homes.

# V. THE DECODE AND ANALYSIS SCHEME

In section IV, we have introduced the data collection scheme which provides personalized local differential privacy for residents' sensitive information. We will propose the decode and analysis scheme in this section.

## *A. High-precision Estimation of Privacy Budgets*

To estimate users' unperturbed Bloom filters, the collector needs to know the probability *p*. Nevertheless, since the value of *p* is the crux to protect the user's privacy budget from others, *p* is unknown to the collector. In most cases, privacy budgets themselves are also sensitive information for users, we have to employ the feature of

Bloom filter to protect users' privacy budgets. If *p* is public to the collector, the privacy budget will be known to the collector, then the user's information will face the risk to privacy. Therefore, we must estimate the value of p for the next work.

We propose a high-precision estimation method of privacy budgets here. Based on Theorem 1, we know that the privacy budget $\epsilon$ is the function of *p* (*q* is uniform to in the whole smart home). So we need to estimate the value of *p* to evaluate $\epsilon$. The collected data from a user are just a private Bloom filter and its corresponding security domain. The collector can only estimate the value of *p* through these two data. The collected private Bloom filter consists of one and zero, and the security domain specifies the different range between the non-private Bloom filter and the private one.

Let *X* be the number of ones in the received Bloom filter and [*s*, *e*] be the security domain, we know that the expectation of *X* is:

$$E[X] = a + (e\text{-}s)\cdot b. \qquad (2)$$

Then, we can deduce the following equation:

$$E[X] = 0.25\cdot(1\text{-}2q)\cdot(e\text{-}s+1)\cdot p^2 + [q\cdot(e\text{-}s+1)\text{-}1]\cdot p + 1.(3)$$

When $q$=0.5, the solution of the above equation is $p$=2($X$-1)/($e$-$s$-1); but when $q\neq0.5$, the solution is

$$p = \frac{-2[(e\text{-}s+1)]+2\sqrt{[(e\text{-}s+1)\cdot q\text{-}1]^2\text{-}(1\text{-}X)\cdot(1\text{-}2q)}}{1\text{-}2q}$$

Equation (3) shows we can make use of *X* to estimate the value of *p*, further find out users' privacy budgets.

Compared to the condition that the collector shares the privacy budget with a user, the estimation strategy inevitably introduces a little noise to the value of *p*. However, the strategy brings two benefits to the collector. One is that makes more users in smart homes be willing to contribute their information for aggregation. Another is what implies residents choose lower privacy levels, further is beneficial to the decoding precision of collected private data.

### *B. High-accuracy Decode*

In order to learn the frequency of each investigated data, the collector needs to estimate the number of ones in its corresponding bit among all original Bloom filters. Assuming that there are *N* respondents in the smart home, and the number of ones in each bit *i* is $O_i$ for all Bloom filters. Mi denotes the accumulation of the *i-th* bit which belongs to its relevant security domain. When every respondent has a common privacy budget in the smart home, the expected number of ones on the *i-th* bit among all collected perturbed Bloom filters is given as follows:

$$E[C_i] = O_i\, a + (M_i - O_i). \qquad (4)$$

Similarly, to estimate the value of Oi, the collector needs to count the number of ones on the *i-th* bit among all private Bloom filters $C_i$ and the value of $M_i$. Then solves (4), we can get the following equation:

$$E[O_i] = \frac{C_i\text{-}M_i b}{a\text{-}b}. \qquad (5)$$

Since every respondent may choose a distinct privacy level, the above equation no longer applies to the scenario we present in this paper. We now have to extend the estimating method in (5) to the case users' privacy budgets are diverse. By means of a little deformation to (5), we get the streaming of form of (5) which implies that the estimation of Oi is an accumulation of ones and zeros in the *i-th* bit which belongs to security domains. Consequently, the estimation of Oi which applies to the personalized situation is given by

$$E[O_i] = \frac{\sum_{k=1}^{C_i}(1\text{-b})\text{-}\sum_{k=C_i+1}^{M_i} b}{a\text{-}b}. \qquad (6)$$

By making use of the streaming form estimate method to each bit of personalized perturbed Bloom filters, we obtain a vector composed by $O_i$, $i\in[1, d]$. The vector is denoted by *V*, and its elements represent the estimated count of ones on each bit among original Bloom filters. During aggregation, the collector applies (6) to incrementally add weight to vector *V* at every received private Bloom filter *B'*.

### *C. The Privacy Preservation Framework for Smart Homes*

In this subsection, we present the whole privacy preservation framework for data collection and analysis. It aggregates residents' personalized private data at the same time preserving their sensitive data and privacy budgets in smart homes.

- *Configure Parameters*: The data collector publishes global parameters to every potential contributor in smart homes before collection. The global parameters include the length of Bloom filters *d* and the probability *q*. Communication between residents and the collector through unsecure networks should be protected. To avoid any adversary to tamper the global parameters maliciously, confidentiality and integrity should be ensured for any data transfer. VPN technique [10] is a commonly used approach for securely transferring sensitive data.
- *Solicit*: The collector sends a query request on a topic of interest to every potential contributor.
- *Randomized Response*: Each resident either declines or replies the query request after he/she receives a query quest and global parameters. When a resident holding a sensitive *m* decides to share his/her sensitive information to the collector, he/she should select a personalized privacy budget to deduce the probability *p* according to Theorem 1. Furthermore, the willing resident still needs to specify his/her preferred security domain [*s*, *e*], then obtains a private Bloom filter *B'* with the use of two-stage binary randomized response algorithm. Finally, the resident sends the collector the perturbed Bloom

filter $B'$ and its corresponding data secure range [$s$, $e$].

- *Collect*: After receiving a response consisting of $B^{'}$ and [$s$, $e$], the collector firstly counts the number of ones in $B'$, denoted by $X$, and makes use of the estimation of privacy budgets algorithm to get the values of $a$ and $b$. Then the collector utilizes (6) to obtain the aggregated vector $V$. $V$ contains the numbers of ones among all collected private Bloom filters on each bit.
- *Decode*: Let $I$ be the set of candidate values among participants in smart homes, and $I_j$ be the $j$-th value of $I$. Be same with decoding process in RAPPOR, the collector firstly constructs a matrix $M$ of size $d \times n$, where $n$ is the size of $I$. $M$ is a sparse matrix where each column is mostly zero with just a one at the Bloom filters for every value in $I$. The key work is to make use of the collected array $V$ and matrix $M$ to infer frequency of each element in interest set $I$. Then the collector applies linear regression technique to fit a model $V \sim M$. The non-zero coefficients are the estimated numbers of their corresponding candidate values of this collection.

## VI. SIMULATIONS AND EVALUATION

In this section, we evaluate the performance of our privacy preservation scheme for smart homes, and verify our proposed data collection and analysis framework over four different synthetic datasets.

The first dataset *norm-age* we use consists of residents' ages that complies with a normal distribution. The *exp-age* dataset contains residents' age information whose frequencies display exponential decay. The last two datasets *unif-t1* and *unif-t2* are all drawn from simulating participants' preference temperatures in smart homes. The two datasets exhibit uniform distribution. The synthetic normal distribution has mean 50 and standard deviation 10, and its corresponding age range is [0, 100]. The simulated exponential distribution has standard deviation 20 with the relevant set of age [0, 100]. The first uniform distribution ranges from [10, 30], and the second is in the interval [20, 30]. The length of Bloom filters we use in all simulations is 128.

Our main goal of this experimental study is to validate that by taking personal privacy budgets and security domains into account, our mechanism can usually get more accurate data analysis results, compared to RAPPOR mechanism proposed in [6], which provides a uniform privacy budget. To that aim, we compare the two methods of privacy preservation, in terms of estimation accuracy, KL divergence under different data distributions.

### A. Accuracy of Estimation

In retrospect, our motivation is to enable an untrusted collector accurately learn users' information distribution. This goal can be precisely measured by accuracy of estimation. The accuracy criteria we use here is mean relative error (MRE). For each actual frequency $F_i$, if its corresponding estimated frequency is $F_i'$ where $i$ belongs to the set $I$, we know the mean relative error is given by:

$$\text{MRE}=\frac{\sum_{i\in I}|F_i^{'}\text{-}F_i|}{\sum_{i\in I}F_i}$$

The simulations are performed with 100,000 residents' reports. For the convenience of experiments, we specify a uniform privacy budget for every participant. The privacy constraints are: $p$=0.5 and $q$=0.5 providing $\epsilon$=2ln(3) differential privacy. In order to compare with PLDP, the simulations on RAPPOR mechanism also provide $\epsilon$ =2ln(3) differential privacy with $p$=1/6, $q$=5/6, $f$=0.25 and one function.

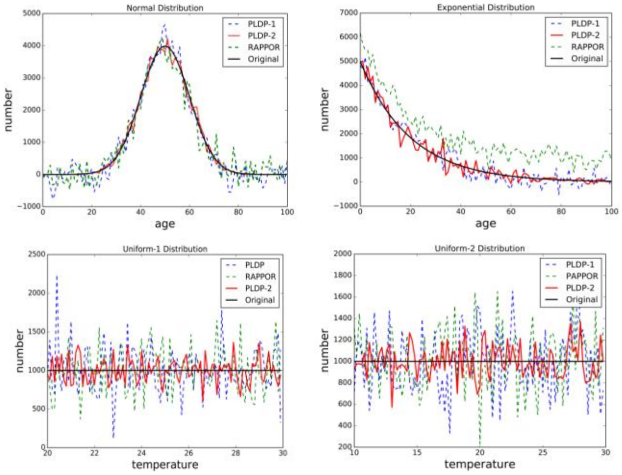

Figure 1. Distribution curves of different schemes.

TABLE I. ACCURACIES OF ESTIMATION UNDER DIVERSE SCHEMES

| **Dataset** | ***norm-age*** | ***exp-age*** | ***unif-t1*** | ***unif-t2*** |
|---|---|---|---|---|
| **PLDP-1** | 0.22352 | 0.22222 | 0.2334 | 0.23482 |
| **RAPPOR** | 0.25187 | 0.92107 | 0.2249 | 0.22662 |
| **PLDP-2** | **0.06766** | **0.16744** | **0.13241** | **0.12281** |

We conduct three simulations on four kinds of datasets severally to compare the performance between the two mechanism comprehensively. The first experiment performs our proposed algorithm without security domain, denoted by PLDP-1 here, that is to say, the security domain is the overall range of candidate values; the second one also performs our PLDP algorithm, but with security domain [$m_i$-$d$, $m_i$+$d$] where $m_i$ is the resident $u_i$'$s$ sensitive data, simply denoted by PLDP-2; the last one we simulate makes use of the RAPPOR algorithm, signified by RAPPOR. For the normal distribution simulations, the users specify their security domains as [$m_i$-5, $m_i$+5]; in the second set of simulations, simply denoted by PLDP-2; the last one we simulate makes use of the RAPPOR algorithm, signified by RAPPOR. For the normal distribution simulations, the users specify their security domains as [$m_i$-5, $m_i$ +5]; in the second set of simulations, the users specify [$m_i$ -10, $m_i$+10] as their security domains; in the last two groups, the users specify [$m_i$-0.5, $m_i$+0.5] as their security domains.

We conduct 16 simulation experiments in total, and these simulations are divided into four groups which are denoted by *norm*, *exp*, *unif*-1 and *unif*-2 respectively. To compare the performance of different approaches, we use Original to denote the original distribution without noise in each group. We give the distribution curves and accuracies of estimation under diverse schemes in Figure 1 and Table I respectively. It can be observed that the personalized local differential privacy with security domain scheme we propose substantially outperforms other two schemes. We highlight the smallest accuracies of estimation in Table I. And the simulation results also show that there is no obvious divergence between PLDP-1 scheme and RAPPOR except the selected dataset is *exp-age*. However, the results of the second group indicate that PLPD scheme exceeds RAPPOR too much when users' data obey exponential distribution.

Then we can conclude that the performance of our personalized scheme is approximate to RAPPOR scheme, although estimating privacy budgets brings much more noise than RAPPOR mechanism. And after applying individual security domain, PLDP obtains reasonably small error for all four datasets. Roughly, the difference between PLDP-1 and PLDP-2 proves the benefit of introducing the notion of security domain. In addition, the proposed scheme also protects users' personalized privacy requirements, and the process of data collection and analysis will benefit from this feature of our scheme

### B. *KL Divergence*

Other than estimating frequencies, we need also to accurately learn residents' information distribution over some domain. KL divergence expresses the discrepancy between estimated distribution and actual distribution. We use it to evaluate the impact of PLDP-2 mechanism on data distribution now. Moreover, to better get a sense of the effect of our approach, we also simulate with RAPPOR mechanism on all the synthetic datasets. The number of contributed residents is 100,000, and the adopted privacy budget is 2ln3 among all simulations. We give the KL divergences of PLDP-2 and RAPPOR over the four datasets in Table II. We emphasize the smaller KL divergences, and find out that PLDP-2 is always superior to RAPPOR. All these results together validate the importance of our personalized local differential privacy model for smart homes.

TABLE II. KL DIVERGENCE UNDER DIVERSE SCHEMES

| Dataset | *norm-age* | *exp-age* | *unif-t1* | *unif-t2* |
|---|---|---|---|---|
| **RAPPOR** | 0.25390 | 0.28128 | 0.039988 | 0.038032 |
| **PLDP-2** | **0.00573** | **0.03576** | **0.013245** | **0.011569** |

### C. *Benefits of Personalized Privacy*

In the last set of experiments, we study the performance of our framework under different privacy budgets and number of participants. The simulations with 100,000 residents are denoted by less, and more indicates the simulations with 200,000 residents in Fig.2.

As can be observed in Fig.2, PLDP-2 is quite effective. In the worst case, it still provides high estimation accuracy where the chosen privacy budget is enough small. We can further observe that in practice the relative error of PLDP-2 roughly decrease linearly with the adopted privacy budgets increasing. Fig.3 also shows that the more cases have higher frequencies estimation accuracies than the less cases. This implies that the collector could obtain more precise results with the number of contributed residents growing.

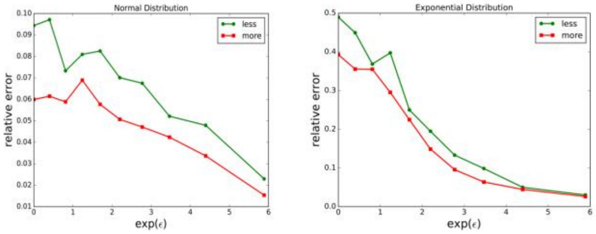

Figure 2. Relative error under different settings.

All the experimental results demonstrate that the avenues of applying personalized local differential privacy with security domain to data privacy cloaking. Using the privacy preservation scheme for smart homes, as a result of being able to specify privacy budget and keep it from others, residents tend to share their sensitive information with the collector. Besides the collector also don't need to choose an enough small privacy budget to satisfy every participant. The collector will benefit from more participants and larger privacy budgets. These results together show that our proposed PLDP-2 scheme is very appropriate for privacy preservation in smart homes.

## VII. CONCLUSION

We have introduced a personalized local privacy-preserving scheme with security domain. The scheme combines the strength of differential privacy with the added flexibility of user-specific privacy levels and security domains. Mechanisms based on Bloom filters and binary response can achieve PLDP-2 effectively and efficiently. The proposed mechanism provides a more desirable trade-off between privacy preservation and data utility for smart homes than RAPPOR mechanism. Furthermore, since custom privacy budgets are potentially sensitive information for contributed residents under local setting, we propose an estimation algorithm to protect them. As the building block scheme. And we believe that our work belongs to an important step to better privacy protection in data collection and analysis for smart homes.

## ACKNOWLEDGMENT

This paper is supported by the National Science Foundation of China under No. U1301256 and 61472385, Special Project on IoT of China NDRC (2012-2766).

## REFERENCES

[1] C. Dwork. Differential privacy: A survey of results. In *TAMC. Springer*, 2008.

[2] Bloom B H. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 1970.

[3] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American S tatisticalAssociation* , 1965.

[4] U. Erlingsson, V. Pihur, and A. Ko rolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS. ACM* , 2014.

[5] S. Wang, L. Huang, M. Tian, et al. Personalized Privacy-preserving Data Aggregation for Histogram Estimation. In *GLOBECOM. IEEE*, 2015.

[6] A. Chakravorty, T. Wlodarczyk T and C. Rong. Privacy Preserving Data Analytics for Smart Homes. In *SPW. IEEE*, 2013.

[7] Rui Chen, Haoran Li, AK Qin, Shiva P Kasiviswanathan, and Hongxia Jin. Private spatial data aggregation in the local setting. In *ICDE. IEEE*, 2016.

[8] G. Drosatos, P. Efraimidis. Privacy-preserving statistical analysis on ubiquitous health data. *8th International Conference on Trust, Privacy and Security in Digital Business*, *Springer-Verlag*, *pp.24-36*, 2011.

[9] S. Moncrieff, S. Venkatesh, et. al. Dynamic Privacy in a Smart House Environment. *IEEE International Conference on Multimedia and Expo*, *pp.2034-2037*, *Jul*. 2007.

[10] CohesiveFT. VPNCubed. *http://www.cohesiveft.com/vpncubed/*, 2008.

[11] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD*, 2006.

[12] C. Dwork. Differential privacy. In *ICALP*. *Springer*, 2006

[13] K. Courtney, G. Demiris, et. al. Needing smart home technologies: the perspectives of older adults in continuing care retirement communities. *Informatics in Primary Care*, *vol.16, pp.195-201*, 2008.

[14] G. Abowd, A. Bobick, et. al. The Aware Home: A living laboratory for technologies for successful aging. *American Association for Artificial Intelligence*, 2002.

[15] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *STOC.ACM*, 2015.

[16] SP Kasiviswanathan, HK Lee, K Nissim, S Raskhodnikova, and A Smith. What can we learn privately? In *FOCS*, 2008.

[17] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *ICML, arXiv preprint arXiv:1602.07387*, 2016.

[18] Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? personalized differential privacy. In *ICDE*. *IEEE*, 2015