Future Generation Computer Systems 109 (2020) 488–499

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Optimization of privacy-utility trade-offs under informational self-determination

Thomas Asikis *, Evangelos Pournaras

Professorship of Computational Social Science ETH Zurich, Zurich, Switzerland

HIGHLIGHTS

- A generic, novel framework for measuring & optimizing privacy-utility trade-offs.
- An analytical proof & application to real-world data from a Smart-Grid pilot project.
- Privacy-utility tradeoffs are optimized under informational self-evaluation.

ARTICLE INFO

Article history:
Received 30 September 2017
Received in revised form 6 May 2018
Accepted 11 July 2018
Available online 18 July 2018

Keywords:
Data sharing
Privacy
Utility
Trade-off
Optimization
Masking
Differential privacy
Data transformation
Diversity
Internet of Things
Big Data

ABSTRACT

The pervasiveness of Internet of Things results in vast volumes of personal data generated by smart devices of users (data producers) such as smart phones, wearables and other embedded sensors. It is a common requirement, especially for Big Data analytics systems, to transfer these large in scale and distributed data to centralized computational systems for analysis. Nevertheless, third parties that run and manage these systems (data consumers) do not always guarantee users' privacy. Their primary interest is to improve utility that is usually a metric related to the performance, costs and the quality of service. There are several techniques that mask user-generated data to ensure privacy, e.g. differential privacy. Setting up a process for masking data, referred to in this paper as a 'privacy setting', decreases on the one hand the utility of data analytics, while, on the other hand, increases privacy. This paper studies parameterizations of privacy settings that regulate the trade-off between maximum utility, minimum privacy and minimum utility, maximum privacy, where utility refers to the accuracy in the estimations of aggregation functions. Privacy settings can be universally applied as system-wide parameterizations and policies (homogeneous data sharing). Nonetheless they can also be applied autonomously by each user or decided under the influence of (monetary) incentives (heterogeneous data sharing). This latter diversity in data sharing by informational self-determination plays a key role on the privacy-utility trajectories as shown in this paper both theoretically and empirically. A generic and novel computational framework is introduced for measuring privacy-utility trade-offs and their Pareto optimization. The framework computes a broad spectrum of such trade-offs that form privacy-utility trajectories under homogeneous and heterogeneous data sharing. The practical use of the framework is experimentally evaluated using real-world data from a Smart Grid pilot project in which energy consumers protect their privacy by regulating the quality of the shared power demand data, while utility companies make accurate estimations of the aggregate load in the network to manage the power grid. Over 20, 000 differential privacy settings are applied to shape the computational trajectories that in turn provide a vast potential for data consumers and producers to participate in viable participatory data sharing systems.

1. Introduction

High data volumes are generated in real-time from users' smart devices such as smartphones, wearables and embedded sensors. Big Data systems process these data, generate information and enable services that support critical sectors of economy, e.g. health, energy, transportation etc. Such systems often rely on centralized servers or cloud computing systems. They are managed by corporate third parties referred to in this paper as *data consumers* who collect the data of users referred to respectively as *data producers*. Data consumers perform data analytics for decision-making and automation of business processes. However, data producers are not always aware of how their data are used and processed. Terms of Use are shown to be limited and ineffective [1,2]. Security and

* Corresponding author.
E-mail addresses: asikist@ethz.ch (T. Asikis), epournaras@ethz.ch (E. Pournaras).

https://doi.org/10.1016/j.future.2018.07.018

privacy of users' data depend entirely on data consumers and as a result misuse of personal information is possible, for instance, discrimination or limited freedom and autonomy by personalized persuasive systems [3–6]. Giving control back to data producers by self-regulating the amount/quality of shared data can limit these threats [7]. Incentivizing the sharing of higher amount/quality of data results in improved quality of service, i.e. higher accuracy in predictions [8–10]. At the same time, data sharing empowers data producers with an economic value to claim.

Several applications do not require storage of the individual data generated by data producers. Instead, data consumers may only require aggregated data. For instance, Smart Grid utility companies compute the total daily power load or the average voltage stability to prevent possible network failures, bottlenecks, predict future power demand, optimize power production and design pricing policies [11,12]. Privacy-preserving masking mechanisms [13], i.e. differential privacy, accurately approximate the actual aggregate values without transmitting the privacy-sensitive individual data of data producers. Masking is a numerical transformation of the sensor values that usually relies on the generation of random noise and is irreversible.[1]

Privacy-preserving masking mechanisms are studied by calculating metrics of privacy q and utility u. The former represents the amount of personal information that a data producer preserves when sharing a masked data value. The latter represents the benefit that a data consumer preserves when using certain masked data for aggregation, e.g. accuracy in data analytics. Literature work [13–15] shows that privacy and utility are negatively correlated, meaning that an increase on one results in decrease on the other. This paper studies the optimization of computational trade-offs between privacy and utility that can be used to model information sharing as supply–demand systems run by computational markets [7,16]. These trade-offs can be measured by the opportunity cost between privacy-preservation and the performance of algorithms operating on masked data, i.e. prediction accuracy. Trade-offs can be made by choosing different parameters for different masking mechanisms each influencing the mean or the variance of the generated noise distributions [13]. Each parameterization results in a pair of privacy and utility values within a trajectory of possible privacy-utility values.

The selection of parameters for masking mechanisms that maximize privacy and utility is studied in this paper as an optimization problem [14,15]. In contrast to related work that exclusively focuses on universal optimal privacy settings (homogeneous data sharing), this paper studies the optimization of privacy-utility trade-offs under diversity in data sharing (heterogeneous data sharing). This is a challenging but more realistic scenario for participatory data sharing systems that allow informational self-determination via a freedom and autonomy in the amount/quality of data shared by each data producer. A novel computational framework is introduced to compute the privacy settings that realize different privacy-utility trade-offs.

The main contributions of this article are the following: (i) The introduction of a generalized, domain-independent, data-driven optimization framework, which selects privacy settings that maximize privacy and utility. (ii) A formal proof on how high utility can be achieved under informational self-determination (heterogeneous data sharing) originated from the diversity in the privacy settings selected by the users. (iii) The introduction of new privacy and utility metrics based on statistical properties of the generated noise. (iv) The introduction of a new masking mechanism. (v) An empirical analysis of privacy-utility trajectories of more than 20, 000 privacy settings computed using real-world data from a Smart Grid pilot project.

This paper is outlined as follows: Section 2 includes related work on privacy masking mechanisms, privacy-utility trade-off as well as privacy-utility maximization problems. Section 3 defines the optimization problem and illustrates the research challenge that this paper tackles. Section 4 introduces the proposed optimization framework. Section 5 outlines the experimental settings on which the proposed framework is tested and evaluated. Section 6 shows the results of the experimental evaluation. Finally, Section 7 concludes this paper and outlines future work.

2. Related work

Several algorithms are proposed to perform data aggregation without transmitting the raw data. The basic idea behind such algorithms is to irreversibly transform[2] the data, so that the original values cannot be estimated. While doing so, some of the properties of the data should be preserved to accurately estimate aggregation functions such as sum, count or multiplication [7,9,13,17,18]. The masking process enables the data producers to control the amount of personal information sent to data consumers. These methods also ensure that the data remain private even when a non-authorized party acquires them, for example in the case of a man-in-the-middle attack.

2.1. Privacy-preserving mechanisms

An overview of privacy-preserving mechanisms is illustrated below:

2.1.1. Perturbative masking mechanisms

Perturbative masking mechanisms allow the data producers to share their data after masking individual values. Each value is perturbed by replacing it with a new value that is usually generated via a process of random noise generation or vector quantization techniques on current and past data values [13]. Some of the most well-known perturbative masking methods are the following:

Additive noise. A privacy-preserving approach is the addition of randomized noise [18–20]. This approach is often used in differential privacy schemes [19]. Differential privacy is ensured when the masking process prohibits the estimation of the real data values, even if the data consumer can utilize previously known data values or the identity of the individual who sends the data [21]. Algorithms that achieve differential privacy rely on the notion that the change of a single element in a database does not affect the probability distribution of the elements in the database [18,20–22]. Furthermore, the removed element cannot be identified when comparing the version of the database before and after the removal. This is achieved by adding a randomly generated noise to each data value. The distribution of the random noise is parameterized and usually is symmetric around 0 and relies on the cancellation of noises with opposite values. Increasing the number of noise values also increases the noise cancellation, since a larger number of opposite values are sampled. This property can be used to combine differential privacy mechanisms in order to ensure privacy while achieving high utility [23]. Statistical aggregation queries on the masked data return an approximate numerical result, which is close to the actual result. Differential privacy can be applied to discrete and continuous variables for the calculation of several aggregation functions [9]. Differential privacy can be combined with the usage of deep neural networks [24,25], to apply

[1] It is computationally infeasible to compute the original data using the transformed data.

[2] A process also known as masking.

more complex aggregation operations on statistical databases. Furthermore, several additive noise implementations are susceptible to noise filtering attacks, such as the use of Kalman filters [26] or reconstruction attacks [27]. These attacks can be prevented when the noise is not autocorrelated or the distribution of its autocorrelation is approximately uniform.

Microaggregation. Microaggregation relies on the replacement of each data value with a representative data value that is derived from the statistical properties of the dataset it belongs to. A well-known application of microaggregation is K-anonymity. K-anonymity relies on the notion that at least K original data values are mapped to the same value [28]. When a crisp clustering algorithm is applied on the data, each data value is mapped to the cluster centroid it belongs to. K is the minimum number of elements in a cluster. Using crisp clustering techniques[3] may result in vulnerabilities to specific attacks, so membership or fuzzy clustering is preferred instead [29]. Membership clustering assigns a data point to multiple clusters with a probability that is often proportional to the distance from each cluster centroid. For membership clustering techniques, usually large amounts of data are required. The storage and computational capacity of sensor devices cannot usually support such processes [13,29].

Synthetic microdata generation. An new dataset is synthesized based on the original data and multiple imputations [13]. The "synthetic" dataset is used instead of the original one for aggregation calculations. The application of synthetic microdata generation on sensor devices may produce prohibitive processing and storage costs. Furthermore, the availability of historical data on each sensor device may not be adequate for such methods to achieve comparable performance and efficiency with the perturbative masking methods [13].

2.1.2. Encryption

Several approaches use encryption to produce an encrypted set of numbers or symbols, known as ciphers. The aggregation operations can be performed on the ciphers and produce an encrypted aggregation value. The encrypted aggregate value can then be decrypted to the original aggregate one, with the usage of the corresponding private and public keys and decryption schemes, providing maximum utility and privacy to the recipient. The encrypted individual values cannot be transformed to the original values without the usage of the appropriate keys from an adversary, so maximum privacy is ensured. Currently, there is extensive research on this area, and there has been a recent breakthrough with the development of fully homomorphic encryption schemes [17,30–32]. Homomorphic encryption schemes though require high computational and communication costs, especially when applied in large scale networks [33,34].

2.1.3. Multi-party computation

Multi-Party Computation (MPC) [35,36] can also be used for privacy-preservation [37] by moving data from one device to another. In such an approach, security and integrity of the data depend on the resilience and security of the network. Most of the methods that rely on encryption can calculate the exact sum of the data, but they can also be violated if an attacker manages to have access to the private key or uses an algorithm that can guess it. Furthermore, in most cases they rely on communication protocols that burden the system with extra computational and communication costs [38]. These costs are often prohibitive for devices such as IoT sensors and smartphone wearables in which computational power and storage are limited [36].

[3] Such as K-Means.

2.2. Privacy and computational markets

A supply–demand system operating on a computational market of data, can be created with the introduction of self-regulatory privacy-preserving information systems [7]. Privacy preservation is utilized to create such systems, for instance by using K-means for microaggregation and different numbers of clusters for each sensor. Varying the number of clusters produces different levels of privacy and utility. The resulting trade-off between privacy and utility is used to create a reward system, where data consumers offer rewards for the data provided by the data producers. The rewards are based on the demand of transformed data that enables the estimation of more accurate aggregate values.

A reward system can be combined with pricing strategies from existing literature on pricing private data [16], in which three actors are introduced: Various pricing functions are proposed to the *Market Maker* so that the privacy-utility of both data consumers and data producers are satisfied. The optimization framework of the current paper can utilize any parametric masking mechanism of the literature mentioned in Section 2.1. The output of the optimization can be used along with pricing functions on participatory computational markets, to create fully functional and self-regulatory data markets.

2.3. Comparison and positioning

The challenge of an automated selection of privacy settings that satisfy different trade-offs is not tackled in the aforementioned mechanisms. Privacy-utility trajectories have not been earlier studied extensively and empirically as in the rest of this paper. The optimization of privacy-utility trade-offs under diversity in data sharing originated from informational self-determination is the challenge tackled in this paper. To the best of the authors' knowledge, this challenge is not the focus of earlier work.

3. Problem definition

Related work [7,8,14,15,39] on privacy-utility trade-offs focuses on the parameter optimization of a single masking mechanism. A masking mechanism is often a noise generation process, which samples random noise values from a laplace distribution and then it aggregates it to the data, for instance the sampled noise is then added to the data to achieve differential privacy [18]. The result of the optimization is usually a vector of parameter values $\theta_{\eta,k}$, for a masking mechanism $\eta$and parameter index k. The pair of the masking mechanism and the parameter values is referred as a *privacy setting* $f_\eta\left(S, \theta_{\eta,k}\right)$ of a set of sensor values $S \in R^1$. This privacy setting produces a pair of privacy-utility values $\hat{q}$, $\hat{u}$, such that:

$$\hat{q} \rightarrow \max(Q) \tag{1}$$

$$\hat{u} \rightarrow \max(U) \tag{2}$$

where $(\hat{q}, \hat{u})$ is a (sub-optimal) privacy-utility pair of values, which is computed by an optimization algorithm that searches for the optimal privacy-utility values pair. $\max(Q)$, $\max(U)$ are the maximum privacy and utility values of a privacy value set Q and a utility value set U. These sets are generated by the application of a masking mechanism.

The optimization of an objective function that satisfies both Relations (1) and (2) simultaneously is an NP-hard problem [15], in the case that privacy and utility are orthogonal ($q \perp u$) or

opposite[4] ($q \uparrow, u \downarrow$), and often intractable to solve, since privacy-utility trade-offs prohibit the satisfaction of both Relations (1) and (2). Particularly, maximizing simultaneously utility and privacy usually yields sub-optimal values, which are lower than the corresponding optimal values computed by optimizing each metric separately [15]. Furthermore, such optimization is applicable for statistical databases [13,21], where data are stored in a centralized system. In such case, a specific privacy setting is chosen by the designer/administrator of the system. As a result, this approach relies on the assumption that a specific privacy setting should be used by all data producers.

However, remaining to a fixed privacy setting may be limited for data producers, especially when a data producer wishes to switch to a different privacy setting to improve privacy further. In this case, the optimization of different objective functions is formalized in the following inequalities:

$$q^* > \hat{q} + \delta \wedge u^* > \hat{u} + c \tag{3}$$

where δ measures the change in privacy, which denotes whether the data producers require higher privacy, $\delta > 0$, or lower privacy $\delta < 0$, from the system. c measures the change in utility, which denotes whether the data consumer demands lower utility, $c > 0$, or higher utility $c < 0$, from the system. Finally, (q^*, u^*) denotes a new (sub-optimal) pair of privacy-utility values, computed by an optimization algorithm that searches for the optimal pair of privacy-utility values with respect to the privacy requirements of data producer and the utility requirements of data consumer expressed by c and δ respectively.

The optimization of an objective function to satisfy Relation (3) is also based on the assumption that all data producers agree to use the same privacy setting. This means that data producers may acquire a different privacy level by changing the value of δ via the collective selection of a different privacy setting. Consequently, a single privacy setting is generated and it produces a pair of privacy-utility values, which satisfy Inequality (3). The value of δ is determined via a collective decision-making process applied by the data producers, e.g. voting between different privacy-utility requirements. Such a system is referred to as a *homogeneous* privacy system, where data producers are able to influence the amount of privacy applied on the data by actively participating in the market, nevertheless they all share the same value for δ. The data consumer can bargain for higher utility by offering higher rewards to the data producers to lower their privacy requirements.

Another challenge that arises is the optimization between privacy and utility when each user decides and self-determines a preferred privacy setting instead of using a universal privacy setting. In such a scenario, inequality (3) is substituted by the following set of inequalities:

$$\begin{aligned} &(q_1^* > \hat{q} + \delta_1) \wedge \cdots \wedge \\ &(q_n^* > \hat{q} + \delta_{|N|}) \wedge (u^* > \hat{u} + c) \end{aligned} \tag{4}$$

where δ_n measures the change in privacy which denotes whether a data producer n belonging to a set of users N requires higher privacy, $\delta_n > 0$, or lower privacy $\delta_n < 0$. q_n^* denotes a new (sub-optimal) privacy value for each data producer n. The value is computed by an optimization algorithm that searches for the optimal privacy value with respect the data producer's privacy requirements expressed by δ_n.

A system in which the inequalities of Relation (4) hold is referred to as an *heterogeneous* privacy system, where each data producer self-determines and autonomously applies a privacy setting based on a preferred privacy value and an expected reward for increasing system utility.

4. Framework

The design of a new privacy preserving optimization framework is introduced in this section to tackle the challenges posed in Section 3. Additive noise masking mechanisms require a lower number of parameters in general and they are often used in privacy-utility optimization [13,15,21]. Each privacy setting is illustrated as an ellipse[5] in Fig. 1a. Each point within the ellipse is a possible privacy-utility pair of values. The ellipse center is chosen based on the privacy and utility mode of the setting. The mode is the value with the highest density. In symmetric distributions, it can be measured via the mean. The vertical radius of the ellipse denotes the dispersion of utility values, while horizontal radius denotes the dispersion of privacy values. Additive noise is stochastic, which means that applying the same privacy setting on the same dataset yields varying privacy-utility values. The choice of an optimal privacy-utility pair cannot be achieved by only evaluating the mode of privacy and utility for each privacy setting. If the privacy-utility values of a privacy setting with high utility mode are varying to a large extend, there is high probability that unexpected non-optimal values are observed. To overcome this challenge, the objective function of the parameter optimization algorithm selects the parameters that minimize the dispersion[6] of privacy-utility values while maximizing the expected utility.

A data producer selects any privacy setting, among different ones, that satisfies personal privacy requirements. The proposed framework divides the range of privacy values in a number of equally sized bins, as illustrated in Fig. 1b. Within each bin, a fitness value is calculated for each privacy setting, based on privacy-utility mode and dispersion Each privacy setting produces privacy values with low dispersion. This is done by applying a lower bound constraint on privacy and utility constraint on the dispersion of privacy values and evaluating only privacy settings that satisfy this constraint, as shown in Fig. 1c. The optimization framework evaluates several privacy settings, to find the parameters that achieve maximum privacy-utility values that vary as little as possible. This is illustrated in Fig. 1d in which the ellipses with the highest utility mode and lowest utility dispersion are filtered for each privacy bin.

In a homogeneous data sharing system, a universal privacy setting is selected by the data producers, via, for instance, voting [41]. Alternatively, in a heterogeneous system, the data producers self-determine the privacy setting independently. Theorem 1 below proves that aggregation functions can be accurately approximated (utility can be maximized) even if different privacy settings from the same of different masking mechanisms are selected.

Theorem 1. *Let the transformation of $|I|$ disjoint subsets of sensor values S_i into the respective subsets of masked values M_i using a certain privacy settings f_i for each such transformation. It holds that the aggregation of the generated multisets of masked values M_i approximates the aggregation of the sensor values multiset S_i:*

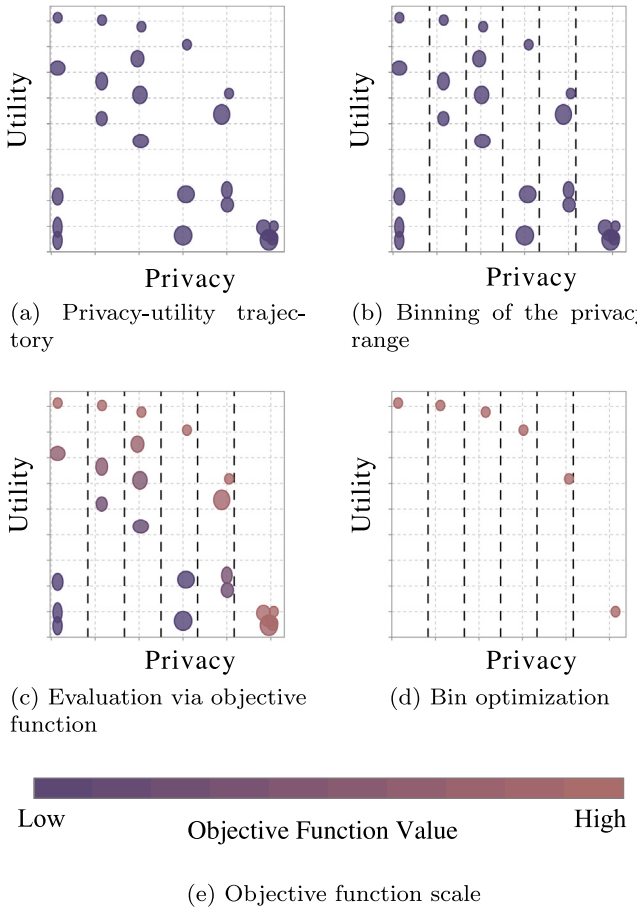
$$g(\bigcup_{i=1}^{|I|} M_i) \rightarrow g(S), \tag{5}$$

given that the commutative and associative properties hold between each of the privacy settings f_i and the aggregation function g.

[4] In the case that privacy and utility are positive correlated ($q \uparrow, u \uparrow$), the problem is reduced to NTIME-hard, and especially in the case privacy and utility are proportional $q \propto u$ to DTIME-hard [40]. The solution of the problem is provided by linearly evaluating all pairs of privacy and utility values once without comparing to all other pairs.

[5] The elliptical shape is chosen for the sake of illustration and it indicates a symmetrical distribution of privacy-utility values, generated by a privacy setting, within the ellipse area.

[6] This refers to the dispersion measures of the privacy and utility distributions. If the values belong to a Gaussian distribution, then the standard deviation is used to measure the dispersion. Since this is not always the case, other measures of scale can be used, such as the Inter-Quantile Range(IQR).



(a) Privacy-utility trajectory

(b) Binning of the privacy range

(c) Evaluation via objective function

(d) Bin optimization

(e) Objective function scale

Fig. 1. A graphical representation of the algorithm. Each ellipse denotes the privacy-utility values of a privacy setting. In Figs. 1c and 1d the varying color denotes the fitness value. A lighter red color denotes higher fitness. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Proof. Let a multiset of real sensor values $S \subseteq R^1$ and $|I|$ disjoint subsets of S such that:

$$\bigcup_{i=1}^{|I|} S_i = S, S_i \neq \emptyset \ \forall i \in \{1, \ldots, |I|\} \tag{6}$$

Let a privacy setting $f : S, \Psi \rightarrow M$ be a pairwise element operation between a set of sensor values $S$and a set of noise values Ψ, that transforms each sensor value $s \in S$ by aggregating it with a randomly selected noise value ψ from Ψ to produce a masked value m:

$$\begin{aligned} &f(S, \Psi) = g(S \cup \Psi) = M \Leftrightarrow \\ &\quad f(s, \psi) = g(\{s, \psi\}) = m \end{aligned} \tag{7}$$

Let $g : A \rightarrow R^1$ be an aggregation function which aggregates all elements of real values multisets S, Ψ, $M \subseteq A \subseteq R^1$ into a single real value $g(A) = z_A \in R^1$. Assume that $g : A \rightarrow R^1$ is defined in a recursive manner so that it satisfies the following equation for a multiset A and any union of all possible combinations of disjoint subsets A_i that satisfy Relation (6):

$$g(A) = g(\bigcup_{i=1}^{|I|} A_i) = g(\bigcup_{i=1}^{|I|} g(A_i)) \tag{8}$$

According to literature [42] the family of aggregation functions that Relation (8) applies to is referred to as extended aggregation functions.[7] The pairwise operation between s and ψ in f is designed in such way that it satisfies the commutative and associative properties when combined with the pairwise operation of g:

$$g(f(S, \Psi)) \overset{(7),(8)}{=} f(g(S), g(\Psi)) \tag{9}$$

where $g(\Psi) \rightarrow \iota$, ι is the strong neutral element of the extended aggregation function g, such that:

$$\begin{aligned} &g(g(A) \cup \iota) = g(A) \Rightarrow \\ &g(g(A) \cup g(\Psi)) \rightarrow g(A) \end{aligned} \tag{10}$$

This property is used in the noise cancellation of Section 2.1.1. Let $|I|$ multisets Ψ_i of noise that satisfy Relation (6), then the following relation holds:

$$\begin{aligned} &g(M_i) = g(f(S_i, \Psi_i)) \overset{(9)}{\Leftrightarrow} \\ &g(M_i) = f(g(S_i), g(\Psi_i)) \overset{(10),(7)}{\Leftrightarrow} \\ &\quad g(M_i) \rightarrow g(S_i), \end{aligned} \tag{11}$$

which means that each noise multiset Ψ_i is generated in such a way that the aggregation of $g(M_i)$ approximates the aggregation of $g(S_i)$. An illustrative example is the laplace noise used in the literature for the aggregation functions of count or summation [18,34], which satisfies Relations (7), (8) and (10). Now it can be proven that:

$$\begin{aligned} &g(\bigcup_{i=1}^{|I|} f_i(S_i, \Psi_i)) \overset{(8)}{=} g(\bigcup_{i=1}^{|I|} g(f_i(S_i, \Psi_i))) \overset{(11)}{\Longleftrightarrow} \\ &\quad g(\bigcup_{i=1}^{|I|} M_i) \rightarrow g(\bigcup_{i=1}^{|I|} g(S_i)) \overset{(6),(8)}{\Longleftrightarrow} \\ &\quad\quad g(\bigcup_{i=1}^{|I|} M_i) \rightarrow g(S) \end{aligned} \tag{12}$$

Thus, Theorem 1 is proven. □

The practical implication of Theorem 1 is that the aggregation of sensor values is approximated by the aggregation of masked values produced by different privacy settings. The approximation stands as long as the noise values produced by the different privacy settings satisfy Relations (9) and (10). According to Relation (6), each subset of sensor values should be masked by one privacy setting. Regarding the complexity of these operations, applying the masking on top of sensor values is linearly depended to the number of sensor values $|S_i|$ assigned to each privacy setting. Due to Relation Eq. (6), applying the proposed framework in real time increases computational complexity by $O(|S|)$. The original values are not stored or transmitted at runtime, thus the storage and communication complexity does not change. During optimization all the privacy settings $i \in I$ are applied to a training set of sensor values S. In that case real sensor values are stored and transmitted as well along with the masked values for each setting. The storage and communication costs increase by $O(|I| \cdot |S|)$. The computation costs also increase to $O(|I| \cdot |S|)$, which is a quadratic complexity in the worst case $|I| = |S|$. In most real world applications, it is safe to assume that the sensor values have considerably higher volume to the evaluated privacy settings $|I| \ll |S|$, thus the expected computational, storage and communication complexity are linear to the number of sensor values.

The framework can be applied as a multi-agent system. It requires two types of agents representing the data consumers and data producers. This scheme can be applied in both centralized and

[7] A subset of those functions are the averaging functions, which include aggregations such as the mean, weighted mean, Gini mean, Bonferroni mean, Choquet integrals etc.

decentralized aggregation services, such as MySQL or DIAS [43]. Finally in both heterogeneous and homogeneous systems, the data consumer can influence the data producer's choice by offering a higher amount of reward to achieve a higher utility.

5. Experimental settings

This section illustrates the experimental settings, which are used to empirically evaluate the proposed framework. A set of sensor values S is used for the evaluation. Each sensor value $s_{n,t}$ belongs to a user n and is generated at time t. For each sensor value, a privacy setting that operates on the device of the data producer masks the sensor value $f_\eta\left(s_{n,t}, \theta_{\eta,k}\right)$ by using the masking mechanism $\eta$with parameters $\theta_{\eta,k}$. Two metrics are used to evaluate privacy and utility.

5.1. Privacy evaluation

The main metric, which is used to calculate privacy, is the difference of the masked value and the original value, which is defined as the local error:

$$\varepsilon_{n,t} = \left| \frac{f_\eta\left(s_{n,t}, \theta_{\eta,k}\right) - s_{n,t}}{s_{n,t}} \right| \tag{13}$$

For a privacy setting to achieve a high privacy, a data consumer should not be able to estimate the local error for the sensor values sent by data producers. This is achieved by choosing privacy settings that generate noise that is difficult to estimate. As it is shown in the literature [7,13,15,21], the noise is difficult to estimate, if it is highly random and causes a significant change in the original value. To avoid noise filtering attacks, noise with low or no autocorrelation is generated. The range of autocorrelation values can be determined analytically when the noise generation function is defined. In case this is not possible, a metric quantifying the color of noise can be included in the objective function. Randomness is evaluated by measuring the Shannon entropy [44] $H(\mathscr{E})$ of the local error for all local error values $\mathscr{E}$. The entropy is calculated by creating a histogram of the error values and then applying the discrete Shannon entropy calculation. Each bin of the histogram has a size of 0.001. The significance of change is measured by calculating the mean local error $\mu(\mathscr{E})$ and standard deviation $\sigma(\mathscr{E})$. When comparing privacy settings, higher mean, variance and entropy indicate higher privacy [13]. In this article, the objective function that measures privacy for a privacy setting $f_{\eta,k}$ is defined as follows:

$$q = \alpha_1 \frac{\mu(\mathscr{E}_{\eta,k})}{\max\left(\mu(\mathscr{E}_{\eta,k})\right)} + \alpha_2 \frac{\sigma(\mathscr{E}_{\eta,k})}{\max\left(\sigma(\mathscr{E}_{\eta,k})\right)} + \alpha_3 \frac{H\left(\mathscr{E}_{\eta,k}\right)}{\max\left(H\left(\mathscr{E}_{\eta,k}\right)\right)} \tag{14}$$

Where $\alpha_1, \alpha_2, \alpha_3$ are weighting parameters used to control the effect of each metric in the privacy objective function. max $(\bullet)$ is the maximum observed value for a metric during the experiments. This value is produced by evaluating all privacy settings $f_{\eta,k}$. Dividing by this value, normalizes the metrics in $[0, 1]$, so that the objective function is not affected by the scale of the metric.

5.2. Utility evaluation

The utility of the system is estimated by measuring the error the system accumulates within a time period, by computing an aggregation function $g(\bullet)$ on the masked sensor values. Examples of such aggregation functions are the daily total, daily average and weekly variance of the sensor values. The accumulated error is referred to as global error[8] and is defined as:

$$\epsilon_t = \left| \frac{g(M_t) - S_t}{g(S_t)} \right| \tag{15}$$

A sample set of global error values ϵ is created by applying the masking process for a number of time periods of the dataset. The mean, entropy and variance of the global error of a privacy setting $f_{\eta,k}$ is calculated over this sample. The mean global error $\mu(\epsilon_{\eta,k})$ indicates the expected error between the masked and actual aggregate. The standard deviation $\sigma(\epsilon_{\eta,k})$ and the entropy $H\left(\epsilon_{\eta,k}\right)$ of the global error, indicate how much and how often the masked aggregate diverges from the expected value. Minimizing all three quantities to 0, ensures that the masked aggregate approximates the actual aggregate efficiently. Thus, after the global error sample is created for each privacy setting, the corresponding utility objective function is calculated:

$$u = 1 - \left(\gamma_1 \frac{\mu(\epsilon_{\eta,k})}{\max\left(\mu(\epsilon_{\eta,k})\right)} + \gamma_2 \frac{\sigma(\epsilon_{\eta,k})}{\max\left(\sigma(\epsilon_{\eta,k})\right)} + \gamma_3 \frac{H\left(\epsilon_{\eta,k}\right)}{H\left(\max\left(\epsilon_{\eta,k}\right)\right)} \right) \tag{16}$$

where the weighting parameters $\gamma_1, \gamma_2, \gamma_3$ are used to control the effect of each metric in the utility objective function. max $(\bullet)$ is the maximum observed value for a metric during the experiments. This value is produced by evaluating all privacy settings $f_{\eta,k}$. Dividing by this value, normalizes the metrics in $[0, 1]$, so that the objective function is not affected by the scale of the metric.

Recall from Section 4 that utility and privacy vary, when repeating the masking process for the same privacy setting and dataset due to the randomness of the noise. A large sample to measure this variance is created, by applying each privacy setting over three times on the same dataset. Then the framework of Section 4 filters the privacy settings based on the mode and the scale of the privacy-utility sample, as illustrated in Fig. 1c. The privacy-utility samples for a privacy setting may not follow a symmetrical or normal distribution.[9] As a result, the maximization of the following objective function is based on utility:

$$\text{perc}(U, 50) + \text{perc}(U, 10) \tag{17}$$

where $perc(U, i)$ calculates the ith percentile of a set of utility values U produced by the application of a privacy setting.

The factors that maximize Relation (17) are: (i) the value of the mode, which is assumed to be approximated by the median and (ii) the dispersion towards values lower than the median, which is expressed by adding the 10th percentile to the median. The objective function evaluates the median and the negative dispersion (10th percentile) of utility values. Positive dispersion is not taken into account in the optimization, since the abstract objective of the optimization is to ensure the least expected utility of a privacy setting for the data consumers. The privacy is constrained by evaluating only privacy settings in which the 10th percentile differs from the privacy median for at most ω, as shown in Inequality (18). The

[8] The error function described in (13) and (15) is also known in literature as absolute percentage error (APE) [45]. The error values are easy to interpret, as APE measures the relative change of the sensor values and aggregate values by using masking. Yet, when the denominator of the function is approaching zero, then the absolute relative error cannot be calculated. If the sensor values are sparse, then another error function can be used, such as MAPE.

[9] It is confirmed in some experimental settings that some privacy settings generate samples of privacy-utility values that do not pass a Kolmogorov Smirnoff normality test [46], and are also non-symmetrical.

value of ω is constrained to be lower or equal to the bin size of the optimization to ensure low privacy dispersion:

$$\text{perc}\,(Q, 50) - \text{perc}\,(Q, 10) < \omega, \tag{18}$$

Where $perc(Q, i)$ calculates the ith percentile of a set of privacy values Q produced by the application of a privacy setting.

6. Experimental evaluation

The proposed framework is evaluated experimentally by applying it to a real-world dataset. Privacy and utility are evaluated using over 20, 000 privacy settings for empirical evaluation.

6.1. Electricity Customer Behavior Trial Dataset

The Electricity Customer Behavior Trial (ECBT) dataset contains sensor data that measure the energy consumption for 6435 energy data producers. The data are sampled every 30 minutes daily for 536 days. For the proposed framework, a set of sensor values $S$of $|N| = 6435$ users and $|T| = 536$ time periods. The total number of sensor values in the set is $|S| = 165, 559, 680$. The sensor data are considered private and the utility company managing the energy network uses them to calculate daily total consumption in the grid, to predict possible failures and plan power production. The daily total consumption is an aggregation that can be defined as the sum of all the sensor values generated during the day: $g\,(S_t) = \sum_{n=1}^{6435} s_{n,t}$. Around 10% of the daily measurements are missing values, and are not included in the experiments. The significance of the missing values reduces as the aggregation interval increases. Therefore, a daily summation is chosen over more granular summation.

During the experiments, the local error of Relation (13) results in a non-finite[10] number only for a low number of maskings. Hence, these values are excluded from the experiments, so that the calculation of finite local error values is feasible. Concluding, the proposed framework operates on 90% of the ECBT dataset.

6.2. Privacy mechanisms

Among several masking mechanisms [13], two ones are used for the evaluation of the framework. Each mechanism is parameterized using the grid search algorithm[11] [47]. The majority of masking mechanisms are parameterized with real numerical values. A grid search discretizes these values, and then evaluates exhaustively all possible combinations of parameter values.

6.2.1. Laplace masking mechanism

This mechanism is widely used in literature [9,13,21]. The noise in the experiments of this paper is generated by sampling a laplace distribution with zero mean. The scale parameter b of the distribution is selected to ensure maximum privacy. Part of privacy can be sacrificed to increase utility if the privacy requirements from the data producers are not high. In this masking mechanism, this is achieved by reducing the b. The scale parameter for each laplace masking setting, is generated from value $b = 0.001$ and during the parameter sweep the value increases by 0.001 until it reaches $b = 10$.

6.2.2. Sine polyonym masking mechanism

This mechanism is introduced in this paper. The mechanism generates random noise that can be added to each sensor value. Assume a uniform random variable υ. The noise generated from the introduced masking mechanism is calculated as follows:

$$m = \sum_{\xi=0}^{|\Xi|} [\theta_\xi \sin(2\pi\upsilon)]^{2\xi+1} \tag{19}$$

The coefficients of the polyonym are denoted as θ_ξ, and ξ denotes the index of the coefficient. Both the length of the polyonym $|\Xi|$ and the individual coefficient values can be tuned to optimize the resulting privacy-utility values of the masking mechanism. The generated noise is symmetrically distributed around zero, because the odd power of the sine function produces both negative and positive noise with equal probability. The sine function and its odd powers are always symmetrical towards the horizontal axis, meaning that $\left|[\theta_\xi \sin(2\pi\upsilon)]^{2\xi+1}\right| = \left|[-\theta_\xi \sin(2\pi\upsilon)]^{2\xi+1}\right|$. Hence, the integral of each factor is zero $\int_0^{1.0} [\theta_\xi \sin(2\pi\upsilon)]^{2\xi+1} d\upsilon = 0$. Therefore the distribution of generated values is symmetrical around zero for $\upsilon \in [0, 1]$, which denotes that the global error mean is approximating zero. Increasing the length of the polyonym and the values of its coefficients, increases the magnitude of the local error, without affecting the global error, indicating that higher utility can be achieved without sacrificing privacy. These properties make polyonyms of trigonometric functions, such as sine and cosine, eligible candidates for additive noise optimization. By increasing the polyonym length and tuning the coefficient values, a larger space of privacy settings is searched to maximize privacy and utility.

Each coefficient is assigned to a value in the space [0, 1.8]. The grid search in that space starts with a step of 0.03 until the value of 0.3, to evaluate settings that create low noise. Then the step changes to 0.3 until the value of 1.8, to evaluate privacy settings that generate higher values of noise. The sine polyonym masking settings are generated by creating all possible permutations of these values for 5 coefficients. This yields around 10,000 masking settings. Preliminary analysis on the autocorrelation and the spectrograms of the proposed sine polyonym noise does not show autocorrelation and recurring patterns over different spectrograms.[12]

6.3. Error analysis

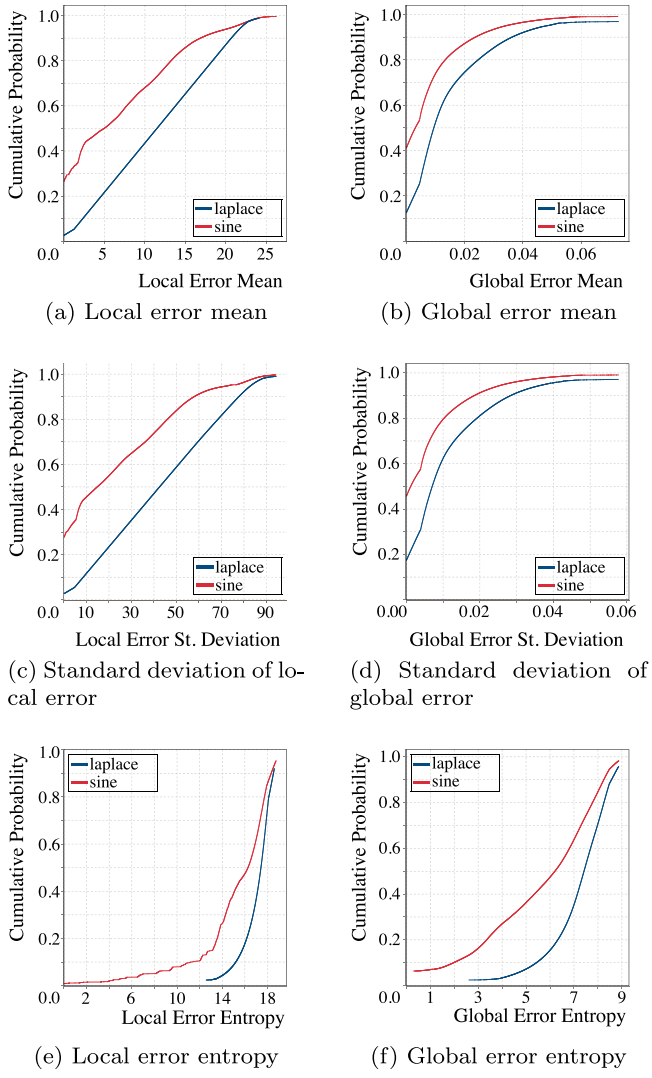
Each privacy setting that results from parameterization of the mechanisms is evaluated by analyzing the local and the global error that they generate on varying subset sizes of the ECBT dataset. By sampling varying sizes of the dataset, the utility and privacy dispersion metrics are evaluated on a varying number of sensor values, measuring the effect of varying participation in the system. To create a random subset of the ECBT dataset, a subset of users N_{test} is chosen. In each repetition the users are chosen randomly. All users use a universal privacy setting. The initial size of the subset is 50 users, and then it increases by 50 users until $|N_{\text{test}}| = 500$ users. Then, the size of the subset increases by 500 users until $|N_{\text{test}}| = 6435$. This process generates several local and global error values. The average, standard deviation and entropy of the local error and global error are calculated for all samples generated from the above process. The empirical cumulative distribution function[13] (CDF) is shown for each metric in Fig. 2.

[10] The original sensor value is zero, therefore the result of Relation (13) is infinite for non-zero noise or indefinite for zero-noise.

[11] Also known as parameter sweep.

[12] Further analysis on this, is possible future work and is out of the scope of this article. This can be evaluated by introducing a metric that measures noise color in the privacy function.

[13] The cumulative distribution function denotes the probability of a generated value being lower or equal than the corresponding domain axis value [48].



(a) Local error mean (b) Global error mean

(c) Standard deviation of local error (d) Standard deviation of global error

(e) Local error entropy (f) Global error entropy

Fig. 2. Cumulative distribution function of each local and global error metric computed by all settings of each masking mechanism.

The sine polyonym mechanism can produce a wider range of local and global error values compared to the laplace mechanism, since almost every sine polyonym CDF curve is covering a wider domain range on the domain axis compared to the respective laplace CDF curves. The majority of the range axis values of the sine polyonym CDF curve are higher than the corresponding range values of the laplace CDF curve. This indicates that it is more probable to generate lower global or local error value by using a sine polyonym setting compared to a laplace setting. Concluding, the sine polyonym settings are expected to produce a wider range of privacy-utility trade-offs. Based on the CDF charts, sine polyonym settings are more likely to achieve higher utility, whereas laplace settings are expected to achieve higher privacy.

6.4. Parameter analysis

For the experiments, α and γ parameters are defined to calculate the privacy and utility. The choice of these parameters may vary based on the distribution of the sensor values and the kind of aggregation. Also data producers and data consumers may have varying requirements that affect the choice of those values. In this paper, these values are determined empirically, to showcase an empirical evaluation. If a data consumer successfully calculates the local error mean by acquiring the corresponding original values of a masked set, then it is possible to estimate the original sensor values of other masked sets as well, by subtracting the calculated mean. This challenge is addressed by using privacy settings with high noise variance. Still, high variance does not guarantee that the masking process is not irreversible. If noise varies between a small finite number of real values, then the data consumer can also estimate the original value of the data by subtracting the variance. To overcome this challenge, privacy settings that produce noise with high entropy, therefore high randomness, are chosen. Consequently, a lower value for the coefficient of local error mean is chosen as $\alpha_1 = 0.2$, while entropy and standard deviation of the local error share a higher coefficient value of $\alpha_2 = \alpha_3 = 0.4$.

Assigning values to the utility coefficients depends highly on the preferences of the data consumer. In the case of sum, the global error mean should be near 0, unless the data consumer estimates the mean and then subtracts it from the aggregation result. For this paper the main concern is to keep a global error mean near zero, to avoid the aforementioned correction process. Standard deviation and entropy are assigned with equal weight. Therefore, a very high coefficient of $\gamma_1 = 0.6$ for the global error mean is chosen, whereas the coefficients of $\gamma_2 = \gamma_3 = 0.2$ for global error, standard deviation and entropy are chosen. To avoid evaluating mechanisms with high utility dispersion and low utility mode values, a hard constraint is applied and only mechanisms that generate mean $\mu(\epsilon) < 0.1$ and standard deviation values $\sigma(\epsilon) < 0.1$ are evaluated. The normalizing factors of Relations (14) and (16) are chosen after the application of this constraint.

A sensitivity analysis of the parameters for each masking mechanism is performed to evaluate the effect of different parameter values on the privacy and utility output of each masking mechanism. In the laplace masking mechanism, increasing the scale parameter b of the distribution, also increases the total noise added to the dataset. In the sine mechanism, increasing the number and values of the coefficients, also increases the total generated noise. In Fig. 3, a comparison of privacy and utility is shown between the two types of mechanisms. The values of utility and privacy are generated as shown in Section 6.3. The total noise is generated by measuring the noise level of each privacy setting on a sample of 100,000 sensor values.[14] The lines are smoothed by applying a moving average, to make the comparison clearer. For the same amount of total absolute generated noise $\sum_t |\psi_t|$, the laplace privacy settings achieve higher privacy, often more than 1% over the sine polyonym privacy settings. The sine polyonym privacy settings achieve higher utility around 1% over the laplace privacy settings. Therefore the results illustrated in Fig. 2 are reflected in the privacy and utility values generated from the above parameterization. Moreover, the trade-off between privacy and utility is observable, as privacy increases with the decrease of utility and vice versa for both mechanisms.

6.5. Homogeneous system evaluation

All the generated privacy settings are evaluated via the framework proposed in Section 4. The proposed framework filters out five privacy settings for five privacy bins of size 0.2. The constraint value for evaluating privacy settings is chosen empirically to be half of the bin size $\omega = 0.1$, to ensure low privacy dispersion, based on Relation (18). The resulting privacy settings are summarized in Table 1. The last two columns of the table, illustrate the median privacy and utility values for each masking mechanism. The first column shows the id of each setting, which is used as reference in Figs. 4 and 5.

[14] This sample size is chosen to be large enough for statistical significance and small enough to reduce computation costs.

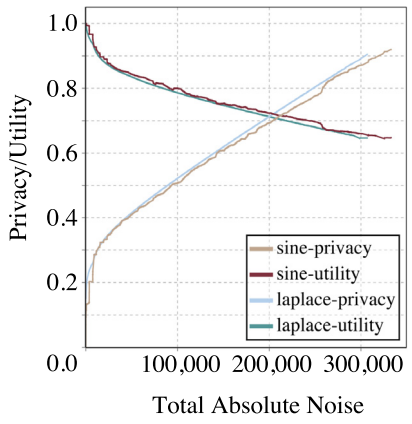
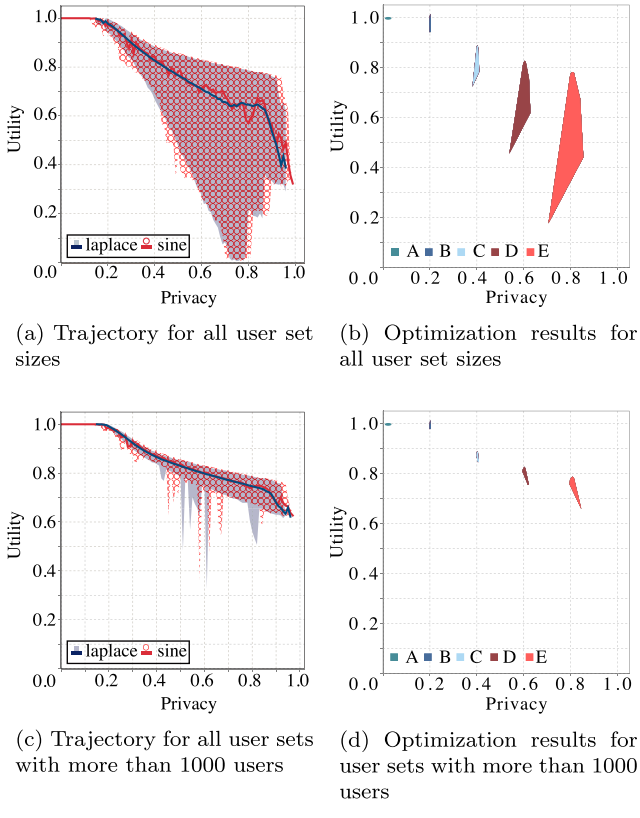


Fig. 3. Comparison of sine polyonym and laplace masking mechanisms in terms of privacy and utility.



(a) Trajectory for all user set sizes

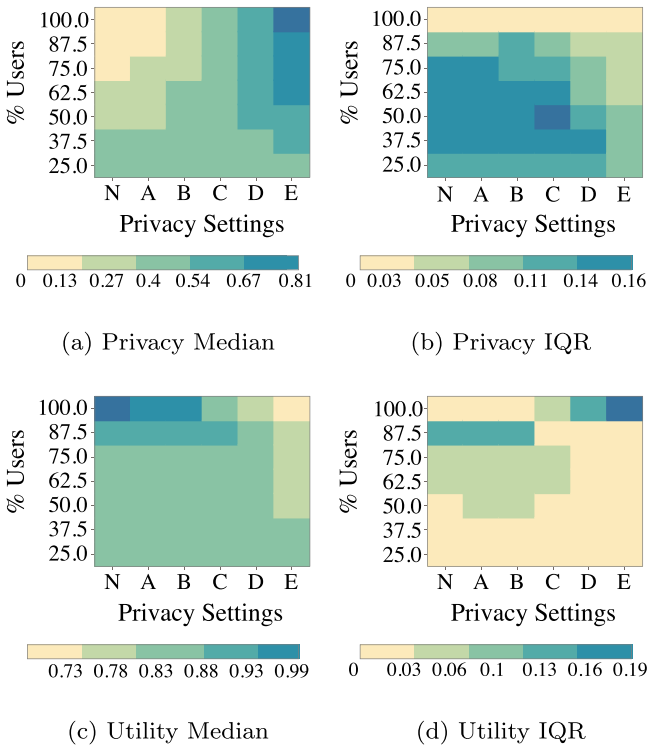
(b) Optimization results for all user set sizes

(c) Trajectory for all user sets with more than 1000 users

(d) Optimization results for user sets with more than 1000 users

Fig. 4. Figs. 4a and 4c show the privacy-utility trajectory of the privacy settings grouped by masking mechanisms in the same color. Figs. 4b and 4d illustrate the trajectories of the privacy settings, which are generated by the proposed framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4a shows the generated privacy-utility values for all the privacy settings tested. Each color is mapped to the masking mechanism that is used to produce this setting. The line denotes the median value of utility at the given privacy value. The non-median privacy-utility values occur in the semi-transparent area. Upper and lower edges of the area denote the minimum and maximum utility value for the corresponding privacy value. Lower utility values for a given privacy point are generated from applications of the privacy setting on small subsets of the ECBT datasets, where $|N| \leq 1000$. The number of sensor values decreases with the number of users. Therefore, the noise cancellation is also reduced, as



(a) Privacy Median

(b) Privacy IQR

(c) Utility Median

(d) Utility IQR

Fig. 5. The heatmaps in Figs. 5a–5d show the privacy and utility median and IQR values, for various distributions of privacy setting choices among the users.

mentioned in Section 2.1.1. Hence, subsets with a lower number of sensor values produce lower utility values. The trade-off between privacy and utility is quantified, since the median curve and the edges of the surrounding area indicate a decrease in utility with the increase of privacy. In Fig. 4b, the area of privacy-utility values of 5 privacy settings produced by the optimization process is shown in Section 5.2. Furthermore the "no masking" privacy setting is also considered, where users choose to use no privacy setting and send the values unmasked.

As it is shown, the privacy values of each privacy setting are within a range of lower than 0.2 privacy. The dispersion of utility increases for privacy settings that achieve higher privacy. The importance of offering more rewards for the usage of higher utility mechanisms is validated, since high dispersion of utility is restrictive for accurate sum calculations by the data consumers. Figs. 4c and 4d illustrate the privacy-utility trajectories for more than 1000 users. It is evident that a data consumer can also increase utility and reduce its dispersion by attracting more users. Higher rewards in general, can also attract more users, so the utility dispersion is expected to decrease even more.

6.6. Heterogeneous system evaluation

In an heterogeneous system, the framework performance is evaluated under the use of different privacy settings from each user. The difference of privacy and utility between homogeneous and heterogeneous systems is quantified. This quantification is done by performing an exhaustive simulation. The simulation combines the ECBT dataset and the six privacy settings in Table 1. Every user of the ECBT dataset is assigned a privacy setting from Table 1. The percentage of users that are assigned each privacy setting is parameterized A histogram with six bins is created. Each bin corresponds to the ID of a specific privacy setting from Table 1. The percentage assigned to a bin denotes the percentage of users

Table 1
A table summarizing the performance of the five optimal privacy settings based on the parameters of the sine polyonym denoting the coefficient value for each factor of the polyonym or the scale value for a laplace mechanism. In case of the sine polyonym, the first number from right is mapped to the first factor ($\xi = 1$) and so on.

ID	Masking	Parameters	Privacy	Utility
A	cosine	0.0-0.0-0.0-0.18-0.0	0.01	0.99
B	laplace	0.005	0.20	0.98
C	cosine	0.6-0.6-0.0-0.9-0.3	0.40	0.84
D	cosine	1.2-0.3-0.6-1.2-0.9	0.60	0.76
E	cosine	1.5-1.5-1.2-0.3-1.2	0.80	0.68
N	None	–	0.00	1.00

using the respective privacy setting at this time point. To generate several possible scenarios for different distributions of user choices, the histogram is parameterized via a parameter sweep of all possible percentage values for each setting, with a step of 12.5%. This process produces over 1000 possible histograms. In Figs. 5a–5d the heatmaps show the median and the interquantile range (IQR)[15] of privacy and utility for all histograms that the privacy setting has a higher percentage of users compared to the others. Such a setting is referred to as dominant setting. This sorting of settings is done to examine the privacy-utility changes while users move from a higher to the next lower utility setting. The top row of the heatmap shows the homogeneous scenario case, where 100% of the users chose only one setting.

The analysis of the heatmap in Fig. 5a shows an increase in privacy when the majority of users choose the more privacy-preserving settings of the homogeneous scenario. This effect is observed for any percentage of users for a dominant setting. A decrease in utility median is confirmed in 5c, when the majority of users shifts from less private to more private settings. The trade-off between privacy and utility is preserved in the heterogeneous scenario, regardless of the percentage of users that choose the dominant setting. Privacy values disperse more in heterogeneous systems, according to Fig. 5b, as the percentage assigned to the dominant setting drops. The dispersion of privacy can reach up to 0.16, which is still lower than the bin range. In terms of utility, the dispersion is much lower on average. There is a dispersion of around 0.1 for high utility mechanisms when they are dominant with 87.5% of users. A possible explanation for this is the reduction of noise cancellation of high privacy settings, due to the low percentage of users choosing them. Concluding, changing from a homogeneous system to a heterogeneous system preserves the trade-off between privacy and utility in the median values. Furthermore, the change to a heterogeneous system increases the dispersion of privacy-utility values for all the mechanisms, so the data consumer should expect the aggregates to be less accurate. Still, utility remains over 0.76 even if the IQR is subtracted from the median, indicating that the aggregate is still approximated even in the heterogeneous case. This validates empirically Theorem 1. In both cases it is efficient for the data consumer to shift user privacy choices to high utility mechanisms by offering them higher rewards. The randomness of the generated noise in an heterogeneous system does not create high variance or high expected global errors. Individual privacy is still preserved for all users and their privacy settings. The individual privacy value does not change between heterogeneous and homogeneous systems, since the privacy-setting choice of one user does not affect the added noise to the sensor values of the other ones.

[15] IQR is considered a robust measure of scale, which is especially used for non-symmetric distributions. It measures the range between the 25th and the 75th quantiles.

7. Conclusion and future work

An optimization framework for the selection of privacy settings is introduced in this paper. The framework computes privacy settings that maximize utility for different values of privacy. This framework can be utilized in privacy-preserving systems that calculate aggregation functions over privatized sensor data. The data producers of such system can self-determine the privacy setting of their choice, since it is guaranteed that it produces the desired privacy with very low deviation. For the data consumer of the system, it is guaranteed that if the data producers are incentivized to use low-privacy settings and high utility settings, the approximated aggregate is highly accurate. Analytical as well as empirical evaluation using over 20, 000 privacy settings and real-world data from a Smart Grid pilot project confirm the viability of participatory data sharing under informational self-determination.

For future work, the proposed framework can be improved by incorporating a machine learning process that computes personalized recommendations of privacy settings to each data producer, by identifying the prior distribution of the sensor data and also the preferences of the data producer. Further empirical evaluations of framework can be performed by implementing other aggregation functions and using different datasets. Finally, an analytical proof that the sine polyonym additive noise is not colored and differentially private can be performed.

Acknowledgment

This work is supported by European Community's H2020 Program under the scheme 'ICT-10-2015 RIA', grant agreement #688364 ASSET: Instant Gratification for Collective Awareness and Sustainable Consumerism'.[16]

Nomenclature

A	A multiset of real values. Any capital letter is treated as a multiset of real values, unless stated otherwise.
$g(A)$	a function that aggregates all elements of a set A into a real value. e.g. for sum: $g_{sum}(A) = \sum_{i=0}^{\|A\|} a_i$
$\mu(A)$	The mean value of all elements of a set, where $a_i \in A$.
m (A)	The median value of all elements of a set, where $a_i \in A$.
max (A)	The maximum value of all elements of a set, where $a_i \in A$.
min (A)	The minimum value of all elements of a set, where $a_i \in A$.
$H(A)$	The Shannon's entropy value for all elements of a set, where $a_i \in A$.
$\hat{a}$	A suboptimal value that approaches an optimal value, e.g. $\hat{a} \rightarrow \max(A)$ or $\hat{a} \rightarrow \min(A)$.
a^*	A new suboptimal value that approaches an existing suboptimal value $\hat{a}$.
n	A user
N	A set of users
t	A time index
$s_{n,t}$	A sensor value generated in time $t$by the user n
η	a masking mechanism, which consists of a parametric algorithm that masks the sensor values of a multiset S.
$\theta_{\eta,k}$	A parameterization $k$for a masking mechanism η.

[16] http://www.asset-consumerism.eu/.

υ	A uniformly distributed variable.
$f_\eta\left(S, \theta_{\eta,k}\right)$	a privacy setting consisting of a masking mechanism ηparameterized with parameters $\theta_{\eta,k}$and operating on a set of sensor values S. It produces a masked set of sensor values $f_\eta\left(S, \theta_{\eta,k}\right) = M$, such that $\lvert S\rvert = \lvert M\rvert$.
Q	A multiset of privacy values.
α_i	A parameter that weights the importance of privacy factors for calculating the privacy values.
δ	A parameter that denotes the amount of privacy that the data producer sacrifices or gains over the existing privacy.
c	A parameter that denotes the amount of utility that a data consumer sacrifices or gains over the existing utility value.
U	A multiset of utility values.
α_i	A parameter that weights the importance of privacy factors for calculating the utility values.
γ_i	A parameter that weights the importance of utility factors for calculating the utility values.

References

[1] J. Bhatia, T.D. Breaux, Towards an information type lexicon for privacy policies, in: 2015 IEEE Eighth International Workshop on Requirements Engineering and Law, RELAW, 2015, pp. 19–24.

[2] A.M. McDonald, L.F. Cranor, The cost of reading privacy policies, J. Law Policy Inf. Soc. 4 (3) (2008) 543–568.

[3] D. Helbing, E. Pournaras, Build digital democracy, Nature 527 (7576) (2015) 33–34.

[4] L. Carmichael, S. Stalla-Bourdillon, S. Staab, Data mining and automated discrimination: A mixed legal/technical perspective, IEEE Intell. Syst. 31 (6) (2016) 51–55.

[5] R. Kitchin, The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences, SAGE Publications, 2014.

[6] S. Gkika, G. Lekakos, Investigating the effectiveness of persuasion strategies on recommender systems, in: 2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization, 2014, pp. 94–97.

[7] E. Pournaras, J. Nikolic, P. Velásquez, M. Trovati, N. Bessis, D. Helbing, Self-regulatory information sharing in participatory social sensing, EPJ Data Sci. 5 (1) (2016) 1–24.

[8] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, D. Megí as, Individual differential privacy: A utility-preserving formulation of differential privacy guarantees, IEEE Trans. Inf. Forensics Secur. 12 (6) (2017) 1418–1429.

[9] Y. DuanYitao, Privacy without noise, in: Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09, ACM Press, New York, New York, USA, 2009, p. 1517.

[10] R.L. Krutz, R.D. Vines, Cloud Security: A Comprehensive Guide to Secure Cloud Computing, Wiley Publishing, 2010.

[11] K. Kursawe, G. Danezis, M. Kohlweiss, Privacy-Friendly Aggregation for the Smart-Grid, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 175–191.

[12] A.C. Burns, R.F. Bush, Marketing research, Pearson, 2014.

[13] C.C. Aggarwal, P.S. Yu (Eds.), Privacy-Preserving Data Mining, in: Advances in Database Systems, vol. 34, Springer US, Boston, MA, 2008.

[14] T. Li, N. Li, On the tradeoff between privacy and utility in data publishing, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, ACM Press, New York, New York, USA, 2009, p. 517.

[15] A. Krause, E. Horvitz, A utility-theoretic approach to privacy and personalization, in: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08, AAAI Press, 2008, pp. 1181–1188.

[16] C. Li, D.Y. Li, G. Miklau, D. Suciu, A theory of pricing private data, ACM Trans. Database Syst. 39 (4) (2014) 1–28.

[17] C. Gentry, Computing arbitrary functions of encrypted data, Commun. ACM 53 (3) (2010) 97–105.

[18] C. Dwork, Differential privacy in: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, 2006, pp. 1–12.

[19] C. Dwork, Differential privacy: A survey of results, in: Theory and Applications of Models of Computation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 1–19.

[20] Z. Wang, K. Fan, J. Zhang, L. Wang, Efficient Algorithm for Privately Releasing Smooth Queries, 2013.

[21] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (2014) 211–407.

[22] Z. Wang, C. Jin, K. Fan, J. Zhang, J. Huang, Y. Zhong, L. Wang, Differentially private data releasing for smooth queries, J. Mach. Learn. Res. 17 (51) (2016) 1–42.

[23] P. Kairouz, S. Oh, P. Viswanath, The composition theorem for differential privacy, IEEE Trans. Inform. Theory 63 (6) (2017) 4037–4049.

[24] R. Shokri, V. Shmatikov, Privacy-preserving deep learning, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15, ACM Press, New York, New York, USA, 2015, pp. 1310–1321.

[25] N. Phan, Y. Wang, X. Wu, D. Dou, Differential privacy preservation for deep auto-encoders: An application of human behavior prediction. in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI-16, 2016, pp. 1309–1316.

[26] J.D. Gibson, B. Koo, S.D. Gray, Filtering of colored noise for speech enhancement and coding, IEEE Trans. Signal Process. 39 (8) (1991) 1732–1742.

[27] C. Dwork, A. Smith, T. Steinke, J. Ullman, Exposed! a survey of attacks on private data, Annu. Rev. Stat. Appl. 4 (1) (2017) 61–84.

[28] P. Samarati, L. Sweeney, Protecting Privacy when Disclosing Information: K-anonymity and its Enforcement through Generalization and Suppression, Technical report, Computer Science Laboratory, SRI Internation, 1998.

[29] J. Nin, J. Herranz, V. Torra, On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. 67 (3) (2008) 399–412.

[30] C. Gentry, A Fully Homomorphic Encryption Scheme (Ph.D. thesis), Stanford University, 2009. crypto.stanford.edu/craig.

[31] C. Gentry, Fully homomorphic encryption using ideal lattices, in: Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, STOC '09, ACM, New York, NY, USA, 2009, pp. 169–178.

[32] C. Gentry, S. Halevi, Implementing Gentry's Fully-Homomorphic Encryption Scheme, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 129–148.

[33] C. Gentry, Fully homomorphic encryption using ideal lattices. in: Proceedings of the 41st annual ACM symposium on Symposium on theory of computing STOC 09, 19(September), 2009, p. 169.

[34] Y. Duan, Differential privacy for sum queries without external noise, in: ACM Conference on Information and Knowledge Management, CIKM, 2009.

[35] A.C. Yao, Protocols for secure computations, in: Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, SFCS '82, IEEE Computer Society, Washington, DC, USA, 1982, pp. 160–164.

[36] S. Bennati, E. Pournaras, Privacy-enhancing aggregation of internet of things data via sensors grouping, Sustain. Cities Soc. 39 (2018) 387–400.

[37] W. Du, M.J. Atallah, Secure multi-party computation problems and their applications, in: Proceedings of the 2001 workshop on New security paradigms - NSPW '01, ACM Press, New York, New York, USA, 2001, p. 13.

[38] M. Prabhakaran, M. Rosulek, Cryptographic Complexity of Multi-Party Computation Problems: Classifications and Separations, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 262–279.

[39] D. Sánchez, J. Domingo-Ferrer, S. Martínez, Improving the utility of differential privacy via univariate microaggregation, in: Lecture Notes in Computer Science, Springer, Cham, 2014, pp. 130–142.

[40] R.V. Book, Comparing complexity classes, J. Comput. System Sci. 9 (2) (1974) 213–229.

[41] H. Nurmi, Comparing Voting Systems, Theory and Decision Library A, Springer Netherlands, 2012.

[42] G. Beliakov, A. Pradera, T. Calvo, Aggregation functions: A guide for practitioners, in: Studies in Fuzziness and Soft Computing, 2007.

[43] E. Pournaras, J. Nikolic, A. Omerzel, D. Helbing, Engineering democratization in internet of things data analytics, in: 2017 IEEE 31st International Conference on Advanced Information Networking and Applications, AINA, 2017, pp. 994–1003.

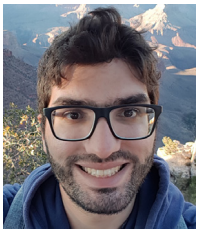
[44] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (July 1928) (1948) 379–423.

[45] S. Makridakis, Accuracy measures: Theoretical and practical concerns, Int. J. Forecast. 9 (4) (1993) 527–529.

[46] P.V. Rao, E.F. Schuster, R.C. Littell, Estimation of shift and center of symmetry based on kolmogorov-smirnov statistics, Ann. Statist. 3 (4) (1975) 862–873.

[47] P.M. Lerman, Fitting segmented regression models by grid search, Appl. Stat. 29 (1) (1980) 77.

[48] M. Spiegel, Schaum's Outline of Statistics, 1992.



Thomas Asikis is a Ph.D. candidate in the Professorship of Computational Social Science, at ETH Zurich, Zurich, Switzerland. Earlier he worked for two years in the industry on the fields of Machine Learning, Telecommunications and Web Development as a data scientist and a full stack developer. He also has a M.sc. in Information Systems and a B.sc. in Management Science. His main fields of interest are privacy-preserving information systems, machine learning and recommender systems. He has a publication in the fields of optimization and recommender systems and is also co-author in a patent under evaluation for the company Incelligent.

Evangelos Pournaras is a senior scientist in the Professorship of Computational Social Science, at ETH Zurich, Zurich, Switzerland. He was earlier at Delft University of Technology and VU University Amsterdam in the Netherlands, where he completed his Ph.D. studies in 2013 with the thesis "Multi-level Reconfigurable Self-organization in Overlay Services". Since 2007, he holds a M.Sc. with distinction in Internet Computing from University of Surrey, UK and since 2006 a B.Sc. on Technology Education and Digital Systems from University of Piraeus, Greece. Evangelos has also been a visiting researcher at EPFL in Switzerland and has industry experience at IBM T.J. Watson Research Center in the USA. He serves the editorial board and the program committees of several international conferences and journals. He has several publications in high-impact journals and conferences in the area of distributed systems, including a best journal paper award. Smart Grids and social sensing/mining are some application domains of his expertise. Evangelos is currently working on the Nervousnet project focusing on fully decentralized and privacy-preserving services for social sensing and mining.