Pournaras et al. *EPJ Data Science* (2016) 5:14
DOI 10.1140/epjds/s13688-016-0074-4

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

Self-regulatory information sharing in participatory social sensing

Evangelos Pournaras[1*], Jovan Nikolic[1], Pablo Velásquez[1], Marcello Trovati[2], Nik Bessis[3] and Dirk Helbing[1]

*Correspondence:
epournaras@ethz.ch
[1] Professorship of Computational Social Science, ETH Zurich, Clausiusstrasse 50, Zurich, 8092, Switzerland
Full list of author information is available at the end of the article

Abstract

Participation in social sensing applications is challenged by privacy threats. Large-scale access to citizens' data allow surveillance and discriminatory actions that may result in segregation phenomena in society. On the contrary are the benefits of accurate computing analytics required for more informed decision-making, more effective policies and regulation of techno-socio-economic systems supported by 'Internet-of Things' technologies. In contrast to earlier work that either focuses on privacy protection or Big Data analytics, this paper proposes a self-regulatory information sharing system that bridges this gap. This is achieved by modeling information sharing as a supply-demand system run by computational markets. On the supply side lie the citizens that make incentivized but self-determined decisions about the level of information they share. On the demand side stand data aggregators that provide rewards to citizens to receive the required data for accurate analytics. The system is empirically evaluated with two real-world datasets from two application domains: (i) Smart Grids and (ii) mobile phone sensing. Experimental results quantify trade-offs between privacy-preservation, accuracy of analytics and costs from the provided rewards under different experimental settings. Findings show a higher privacy-preservation that depends on the number of participating citizens and the type of data summarized. Moreover, analytics with summarization data tolerate high local errors without a significant influence on the global accuracy. In other words, local errors cancel out. Rewards can be optimized to be fair so that citizens with more significant sharing of information receive higher rewards. All these findings motivate a new paradigm of truly decentralized and ethical data analytics.

Keywords: privacy; summarization; analytics; aggregation; self-regulation; social sensing; supply-demand; reward; incentive

1 Introduction

The introduction of 'Internet of Things' has brought paramount opportunities for information sharing in society. 'Big Data' technologies can efficiently process large-scale streams of data generated by mobile phone sensors, embedded systems for smart city infrastructures and smart sensors running residential applications, such as energy management, and ambient assisting living. These new technological opportunities have several implications. On the one hand, accurate data analytics with fine-grained data have the potential for more cost-effective decision-making, policies and regulation systems [1–3]. On the other hand, information sharing challenges privacy of citizens and creates opportunities for massive surveillance and discriminatory actions that result in segregation phenomena in society [4–6]. There is active ongoing work on either improving the accuracy of data analytics [7, 8] or introducing countermeasures for privacy preservation [9, 10]. However, in most cases, these research efforts are canceling out each other. A more effective data analytics methodology requires a lower level of privacy preservation. Similarly, a privacy protection mechanism often deprives data from the analytics process whose accuracy could be critical for the society. This paper bridges this gap by introducing a self-regulatory information sharing system for participatory social sensing.

Information sharing and data aggregation are modeled as a supply-demand system running via a computational market. Citizens who share information are in the supply-side and data aggregators who perform the analytics are in the demand-side. In this context, matching supply and demand requires citizens making incentivized but self-determined choices about a level of information sharing that results in an equilibrium between privacy preservation and accurate analytics. Monetary rewards play a crucial role in incentivizing information sharing [11]. This paper engages such rewards to regulate the equilibrium trade-offs: (i) For citizens, a high privacy preservation comes along with low rewards and a low privacy preservation with high rewards. (ii) For data aggregators, a high accuracy in the analytics results in high costs for the provided rewards. On the contrary, low accuracy reduces the costs. This paper contributes metrics that can quantify these trade-offs. Moreover, the trade-offs of information sharing can be regulated with a generic reconfigurable summarization mechanism that locally controls for each citizen the information reveal to the data aggregators according to his/her privacy and rewards preferences.

The proposed system is generic as it can be applied in several social sensing systems with different types of sensor data. This paper studies information sharing empirically with real-world data from two application domains: (i) Smart Grids with electricity consumption data from more than 6,000 individuals and (ii) mobile phone sensing from several different sensors. Results show striking trade-offs between privacy-preservation, accuracy of analytics and costs from the provided rewards under different experimental settings. The influence of the number of participating citizens, the type of data summarized and the policies under which rewards are distributed among citizens are quantified in several experiments performed. These measurements provide invaluable insights about how to shape the future of self-regulatory information sharing with a truly decentralized and more ethical paradigm for data analytics.

This paper is outlined as follows: Section 2 models information sharing as a supply-demand system and introduces the metrics that govern the equilibrium of such a system. Section 3 illustrates a realization of the proposed system and the experimental methodology for its empirical evaluation. Section 4 illustrates the evaluation of the proposed self-regulatory information sharing system. Section 5 makes comparisons with related work. Finally, Section 6 concludes this paper and outlines future work.

2 A self-regulatory information sharing system

This paper studies participatory social sensing as a decentralized supply-demand system operating over a computational market. The resource traded within this market is high granularity time series data generated by a crowd of participatory citizens. The data are generated with 'Internet of Things' technologies and are a result of some social or environmental activity in an application domain such as energy management [2, 12], traffic management [13], ambient assisting living [14] and other. The supply-side concerns citizens

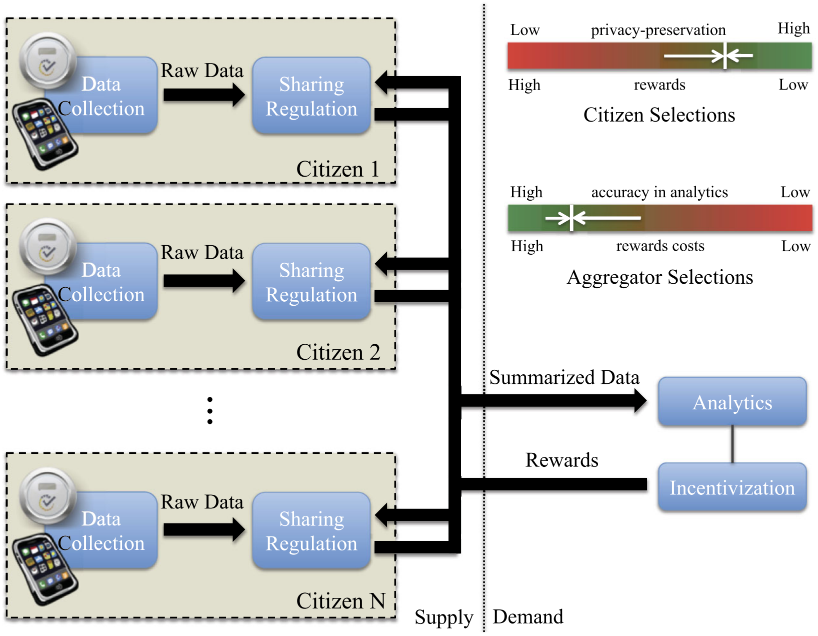


Figure 1 **Privacy-preserving participatory social sensing as a supply-demand system.**

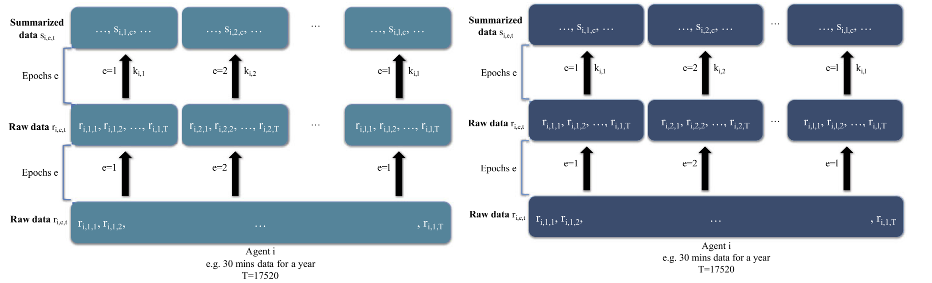
that participate in trading a share of the data they collect. The amount of data they share is governed by two opposing objectives: privacy-preservation of their generated data vs. maximum rewards received for the data they share. On the other hand, demand-side refers to data aggregators that collect data shared by incentivized citizens in order to perform analytics services. Data aggregators also have two opposing objectives: maximum accuracy in analytics vs. minimum costs from the rewards provided to citizens for the shared data. Figure 1 shows a graphical illustration of the supply-demand system envisioned.

Table 1 outlines the mathematical symbols of this paper in the order they appear. The supply-demand system is formally modeled as follows. Let for each citizen a software agent i managing the high granularity time-series data $R_{i,e} = (r_{i,e,t})_{t=1}^{T}$ collected for an epoch e, *e.g.*, a year. This data is referred to as *raw data*. The agent also represents the citizens' preferences of data sharing and can make automated decisions on behalf of the citizen based on these preferences.

Data sharing is regulated by turning the raw data $R_{i,e} = (r_{i,e,t})_{t=1}^{T}$ to the *summarized data* $S_{i,e} = (s_{i,e,t})_{t=1}^{T}$ using a summarization function $f_s(R_{i,e}) = S_{i,e}$, subject to $|R_{i,e}| = |S_{i,e}|$. A summarization function maps the raw data to a limited domain $(c_{i,e,j})_{j=1}^{k_{i,e}}$ of $k_{i,e}$ possible discrete values, nevertheless, the summarized data can represent with a level of uncertainty the raw data so that the analytics performed in the demand-side are accurate to a required level. This provides a more effective privacy protection level and encourages citizens to participate in social sensing platforms. The summarization function is exclusively and privately selected by citizens and can differ among them. This limits the inference opportunities by data aggregators. Summarization can be performed multiple times over the period T by splitting the raw data to a number of l epochs. For example, data collected at every half an hour for a whole year, can be split into daily or weekly epochs on which summarization can be separately applied in a periodic fashion. Figure 2 illustrates this example graphically.

Table 1 An overview of the mathematical symbols

Symbol	Interpretation
i	An agent index
e	An epoch index
t	A time index within an epoch
T	Epoch duration
$R_{i,e}$	Sequence of raw data
$r_{i,e,t}$	A record of raw data
$S_{i,e}$	Sequence of summarized data
$s_{i,e,t}$	A record of summarized data
$f_s()$	Summarization function
j	An index for a possible summarization value
$c_{i,e,j}$	A possible summarization value
$k_{i,e}$	The number of possible summarization values
l	Number of epochs
$\alpha_{i,e}$	Summarization metric
$D_{i,e}$	Sequence of raw or summarization data
$H(D_{i,e})$	Entropy
$p_{i,e,j}$	Probability of a possible value occurring in an epoch
n_t	Occurrence or not of possible value at time t
$\beta_{i,e}$	Diversity
m_t	Change or not between two consecutive time periods t and $t+1$
$\epsilon_{i,e,t}$	Local error
$\varepsilon_{i,e}t$	Global error
n	Number of participating citizens
$\epsilon_{e,t}$	Average local error among citizens
γ_e	Total rewards that data aggregators are willing to provide
$P_r()$	Probability density function for rewards
z	Number of discrete participation levels
$P_s()$	Probability density function for summarization
$\gamma_{i,e}$	Rewards provided to agent i


Summarized data $s_{i,e,t}$
…, $s_{i,1,c}$, …
…, $s_{i,2,c}$, …
…
…, $s_{i,l,c}$, …
Epochs e
e=1
$k_{i,1}$
e=2
$k_{i,2}$
e=l
$k_{i,l}$
Raw data $r_{i,e,t}$
$r_{i,1,1}, r_{i,1,2}, …, r_{i,1,T}$
$r_{i,2,1}, r_{i,2,2}, …, r_{i,2,T}$
$r_{i,l,1}, r_{i,l,2}, …, r_{i,l,T}$
Epochs e
e=1
e=2
e=l
Raw data $r_{i,e,t}$
$r_{i,1,1}, r_{i,1,2},$ … $, r_{i,1,T}$
Agent i
e.g. 30 mins data for a year
T=17520

(a) Daily, $T = 48, l = 365$ **(b)** Weekly, $T = 336, l = 52$

Figure 2 Periodic data summarization at two different granularity levels.

The length of the epoch determines the data that the citizen protects. For example, a daily summarization protects the privacy of data within each day but not across days. The latter would require weekly or monthly summarization in the context of this work. The summarization $\alpha_{i,e}$ of an agent i at an epoch e can be measured as follows:

$$\alpha_{i,e} = 1 - \frac{k_{i,e}}{T}, \tag{1}$$

where $k_{i,e}$ is the number of possible discrete values used to summarize the raw data divided by the total number of measurements that can be observed within the duration of an epoch. The information reveal within the raw data and the summarized data can be measured with the entropy of Shannon's information theory [15]:

$$H(D_{i,e}) = -\sum_{j=1}^{k_{i,e}} p_{i,e,j} \log_2 p_{i,e,j}, \tag{2}$$

where the input data $D_{i,e} = (d_{i,e,t})_{t=1}^{T}$ can be either the raw data such that $D_{i,e} \equiv R_{i,e}$, or the summarized data such that $D_{i,e} \equiv S_{i,e}$. The probability $p_{i,e,j}$ is measured as follows:

$$p_{i,e,j} = \frac{1}{T}\sum_{t=1}^{T} n_t, \quad n_t = \begin{cases} 1 & \text{if } c_{i,e,j} = d_{i,e,t}, \\ 0 & \text{if } c_{i,e,j} \neq d_{i,e,t}, \end{cases} \tag{3}$$

where n_t is the number of occurrences of $c_{i,e,j}$ in the data $D_{i,e}$. Finally, diversity is another notion of information reveal that measures the rate of changes in sensor values occurring within an epoch. It is measured as follows:

$$\beta_{i,e} = \frac{1}{T-1}\sum_{t=1}^{T-1} m_t, \quad m_t = \begin{cases} 1 & \text{if } d_{i,e,t} = d_{i,e,t+1}, \\ 0 & \text{if } d_{i,e,t} \neq d_{i,e,t+1}, \end{cases} \tag{4}$$

where m_t counts whether a change occurs between two consecutive time periods $d_{i,e,t}$ and $d_{i,e,t+1}$. The information loss between raw data and summarized data can be measured with the relative approximation error as follows:

$$\epsilon_{i,e,t} = \frac{|r_{i,e,t} - s_{i,e,t}|}{|r_{i,e,t}|} \tag{5}$$

The data aggregators perform analytics using the summarization data instead of the raw data. This paper studies aggregation functions as a common analytics operation, *e.g.*, summation, average *etc.* An aggregation function provides collective information about the individual measurements performed by citizens. The main challenge for data aggregators is if the aggregation functions can be accurately computed using the summarization data instead of the raw data. The error of an aggregation function, such as the summation, is computed as follows:

$$\varepsilon_{e,t} = \frac{|\sum_{i=1}^{n} r_{i,e,t} - \sum_{i=1}^{n} s_{i,e,t}|}{|\sum_{i=1}^{n} r_{i,e,t}|}, \tag{6}$$

where n is the number of participating citizens. To distinguish the two errors, the $\epsilon_{i,e,t}$ computed by each agent i is referred to as *local error*, in contrast to the *global error* $\varepsilon_{e,t}$ computed by data aggregators. Given that the two metrics are relative, the global error can be compared to the average local error among the citizens. The latter is measured as follows:

$$\epsilon_{e,t} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_{i,e,t} \tag{7}$$

Data aggregators incentivize citizens to share data as follows. Assume that data aggregators have a budget γ_e at epoch e that can use to incentivize and reward citizens to share data. The budget cannot be equally divided among citizens as each citizen may select a different summarization level. In other words, the average reward per citizen should be scaled up or down according to the summarization level selected. This can be achieved with a probability density function $P_{\mathrm{r}}(\alpha_{i,e})$ to incentivize citizens to share more or less data. Given a constant budget γ_e, the $P_{\mathrm{r}}(\alpha_{i,e})$ is continuously updated as follows: When the global error is very high, lower summarization is required and therefore higher rewards can be distributed to low summarization values. In contrast, when a higher global error can be tolerated, higher summarization can be tolerated as well, resulting in a relative increment of the rewards to high summarization values. The market equilibrium can be further studied with mechanism design and game theoretic approaches [16].

The rewards received by each citizen i depends on their selection of a summarization level. The summarization level is a technical concept that citizens may not easily perceive so that a meaningful selection is performed for them. This barrier may become apparent when citizens use and interact with mobile phones to generate data. In such cases citizens can easier select a participation level determined within a range $[1,z]$ of $z \leq k_{i,e}$ discrete options. This approach is documented in related work [17, 18] and is the practice of segmented control recommended in the software engineering of mobile applications.[a] Option 1 corresponds to high rewards but low privacy-preservation, whereas, option z corresponds to low rewards but high privacy-preservation. Selections can be made offline via survey questions or online via interactions with the software agent [19, 20]. Selections made are mapped to the range of summarization values determined by the $k_{i,e}$ possible summarization values.[b] A probability density function $P_{\mathrm{s}}(\alpha_{i,e})$ can be constructed that measures the probability of a user to have a certain summarization level $\alpha_{i,e}$.

Given the total number of citizens n, the total budget for rewards γ_e at epoch e, the probability density function $P_{\mathrm{r}}()$ for the distribution of rewards and the probability density function $P_{\mathrm{s}}()$ for the distribution of citizens' selections of a summarization level, the rewards of a citizen i with summarization $\alpha_{i,e}$ at epoch e are measured as follows:

$$\gamma_{i,e} = \frac{\gamma_e * P_{\mathrm{r}}(\alpha_{i,e})}{n * P_{\mathrm{s}}(\alpha_{i,e})}. \tag{8}$$

The following section illustrates how this model can be empirically used and evaluated with real-world data.

3 Experimental methodology

The proposed self-regulatory information sharing system is evaluated empirically using data from two social sensing projects:

- *The Electricity Customer Behavior Trial - ECBT*: This is a Smart Grid project[c] that studies the impact on electricity consumption of residential and enterprise consumers in Ireland. The project ran during the period 2009-2010 with data from 6,435 participating consumers. Consumption data are collected from smart meters every 30 minutes. Data are pre-processed to daily and weekly epochs and cleaned up to include 68.42% of the original data that correspond to 52 weeks with users having at least 95% of data availability. A 0.19% of missing values are interpolated by the earliest meter read and first following one. Summarization is performed in each daily or weekly epoch.

- *The Planetary Nervous System - Nervousnet*: This is a decentralized smart phone platform for social sensing services [21]. A prototype of the platform was deployed on December 2014 at the 31c3 Chaos Communication Congress in Hamburg, Germany. Data of maximum 12 phone sensors were collected with frequency varying from 30 seconds to 15 minutes from a maximum of 154 users. This paper illustrates results for the following sensors: (i) accelerometer,[d] (ii) battery, (iii) light and (iv) noise. The quality of data generated from mobile phone platforms is a challenge. Data quality is improved by filtering out users with a large proportion of missing values. Moreover, data are normalized to $l = 4$ epochs that correspond to the conference days.

The summarization technique adopted is clustering [22]. It is an unsupervised machine learning technique and it is highly customizable. It has a plethora of implemented algorithms that can be used in different data types and application scenarios. Each raw value is replaced by the centroid with the lowest Euclidean distance. A higher summarization is achieved with lower number of clusters. Other techniques more robust to statistical inference could be used in the future as well. The goal of the experimental evaluation is to measure system performance as follows:

- *Privacy*: Privacy-preservation is measured by averaging the entropy and diversity of the shared information over the total time period.
- *Accuracy*: The accuracy of two aggregation functions, the summation and average,[e] is measured with the global error between the raw and summarization data. Given that both local and global errors are relative metrics, the relation of the local error with the global error can show how the local citizens' selections of a summarization level affect the global outcome of the accuracy in the summation and average.
- *Costs*: This is the amount of rewards provided by the data aggregators to each citizen given a total budget, the summarization selections of citizens and the distribution of rewards among different levels of summarization.

Comparisons are made by varying the following factors:

- *Epoch length*: Datasets can be split into different epochs, *e.g.* daily and/or weekly epochs.
- *Summarization level*: It varies according to the following schemes: (i) different fixed summarization levels for all citizens, (ii) empirically by analyzing relevant survey questions or (iii) algorithmically by using the algorithm of expectation minimization to automatically detect the number of clusters.
- *Distribution of rewards*: Incentivization can be tailored to different citizens' groups by adjusting the distribution of rewards among different selections of summarization levels. A linear, an observational and two optimized distributions of rewards are studied.
- *Number of citizens*: System performance can be evaluated with varying percentage of participating citizens in information sharing.
- *Sensor type*: The values of different sensor types may vary significantly and result in different trajectories for the performance measurements introduced in this paper.

Table 2 outlines the experiments that can be performed with each of the two datasets. For example, the ECBT dataset allows empirical selections of the summarization level and calculation of rewards, whereas the Nervousnet dataset does not. This is because of the survey questions available only in the former dataset. Moreover, ECBT includes a single sensor type, the smart meter, whereas Nervousnet includes several smart phone sensors.

Table 2 An outline of the experiments performed with each dataset

Measurements & variables	ECBT	Nervousnet
Privacy	✓	✓
Accuracy	✓	✓
Costs & Rewards	✓	X
Epoch length	daily & weekly	daily
Summarization level	fixed, empirical & algorithmic	fixed & algorithmic
Number of citizens	✓	✓
Several sensor types	X	✓
Analytics	summation	average

Figure 3 Fixed summarization values and the corresponding number of clusters for daily and weekly epochs.

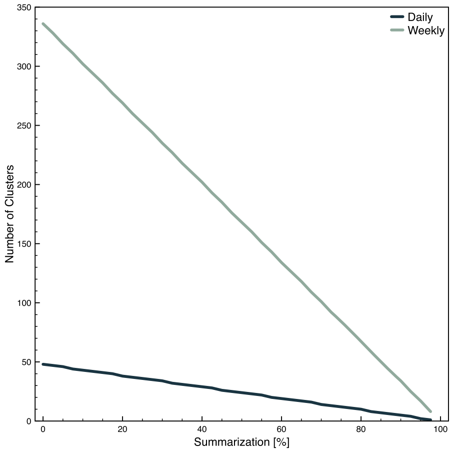


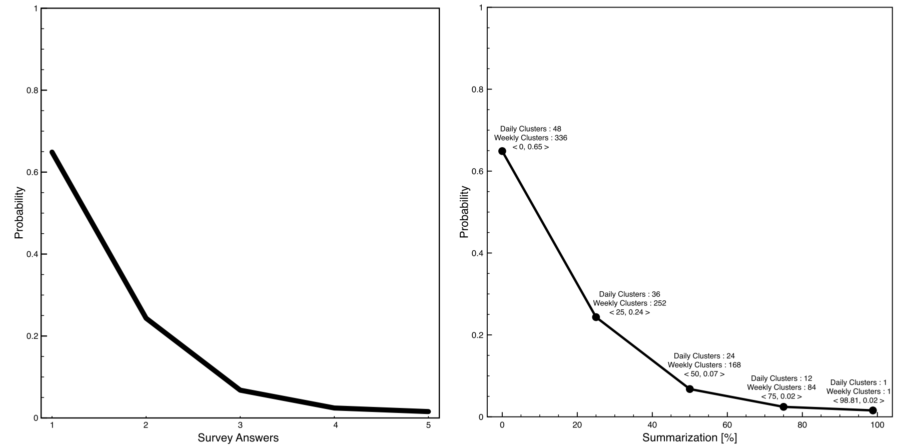
Figure 3 illustrates the scheme with fixed summarization levels. The number of clusters vs. the summarization values are computed with Equation (1).

The empirical selection of summarization levels is performed using the answers of survey questions from the ECBT project. However, the proposed model is generic and can be applied beyond the ECBT project. The goal of engaging these empirical data is to show how the proposed model can be applied in reality rather than studying the actual privacy profiles of citizens. The latter requires highly contextualized data that is challenging to acquire. This expansion is beyond the scope of this paper and is subject of future work. Questions that indicate desire, belief, or intention for a participation in the ECBT project are correlated to the selected level of summarization. The questions have $z = 5$ possible answers with '1' indicating a strong agreement and '5' a strong disagreement. Different questions are answered by residential consumers and small-medium enterprises. For the residential consumers, the following question is considered:

Question 1 My household may decide to be more aware of the amount of electricity used by appliances we own or buy.

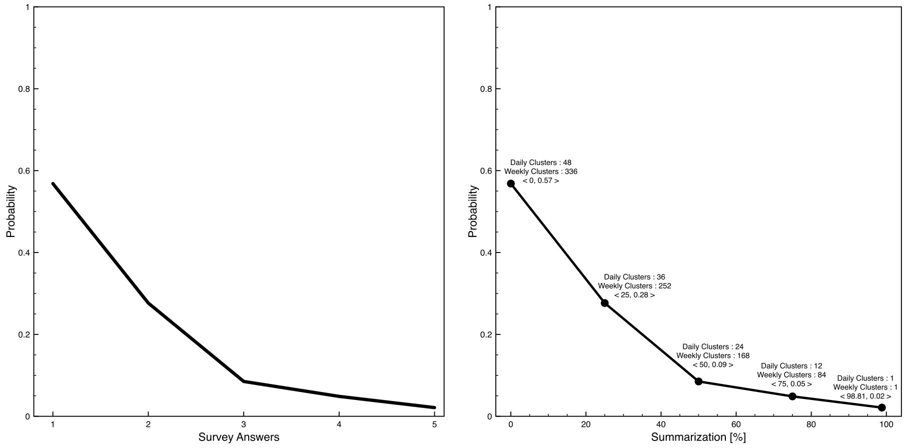
A non-linear exponential half-life regression model approximates[f] the $P_{\mathrm{s}}(\alpha_{i,e})$ from the probability density function of the answers:

$$P_{\mathrm{s}}(\alpha_{i,e}) = a + \frac{b}{2^{\frac{\alpha_{i,e}}{c}}}, \tag{9}$$



(a) (b)

Figure 4 Probability density functions: Survey answers of residential consumers are mapped to summarization values.



(a) (b)

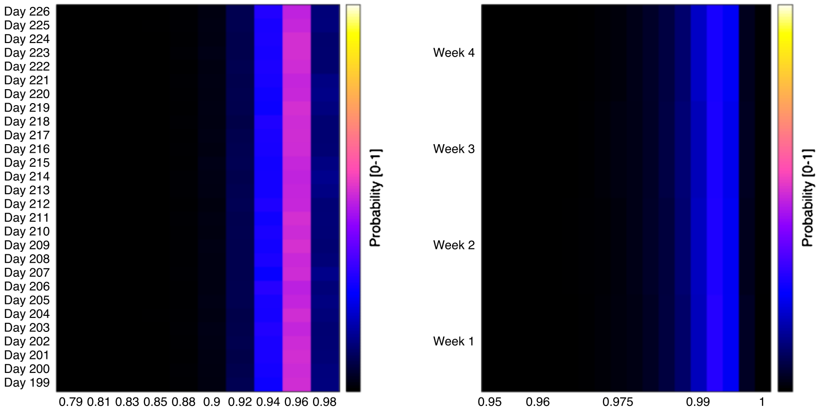
Figure 5 Probability density functions: Survey answers of small-medium enterprises are mapped to summarization values.

where $a = -0.005126568 \pm 0.01334$, $b = 0.655628 \pm 0.01806$ and $c = 17.17108 \pm 1.303$. Figure 4 illustrates the answers of the residential consumers and how these answers are mapped to the probability density function $P_s()$ for the whole range of summarization values. The number of clusters can be derived by solving Equation (1) for $k_{i,e}$.

For the small-medium enterprises, the following question is considered:[g]

Question 2 My organization would like to do more to reduce electricity usage.

The non-linear exponential half-life regression model of Equation (9) also approximates[h] the $P_s(\alpha_{i,e})$ from the probability density function of the answers for $a = -0.0139374 \pm 0.0321$, $b = 0.5861805 \pm 0.03668$ and $c = 22.51562 \pm 3.774$. Figure 5 illustrates the answers of the small-medium enterprises and how these answers are mapped



(a) Daily Summarization

(b) Weekly Summarization

Figure 6 Probability density functions of summarization by automatically detecting the number of clusters in ECBT.

to the probability density function $P_{\mathrm{s}}()$ for the whole range of summarization values. The number of clusters can be derived by solving Equation (1) for $k_{i,e}$.

For the total 30% of residents or enterprises that do not have answers to the questions, a random summarization value is assigned to them according to the probability density function of Figure 4(b).

The algorithmic selection of summarization levels is performed with the algorithm of expectation minimization using 500 iterations and a minimum standard deviation of 10^{-4}. The algorithm automatically detects the number of clusters for each citizen. The probability density function of summarization for ECBT is shown in Figure 6.

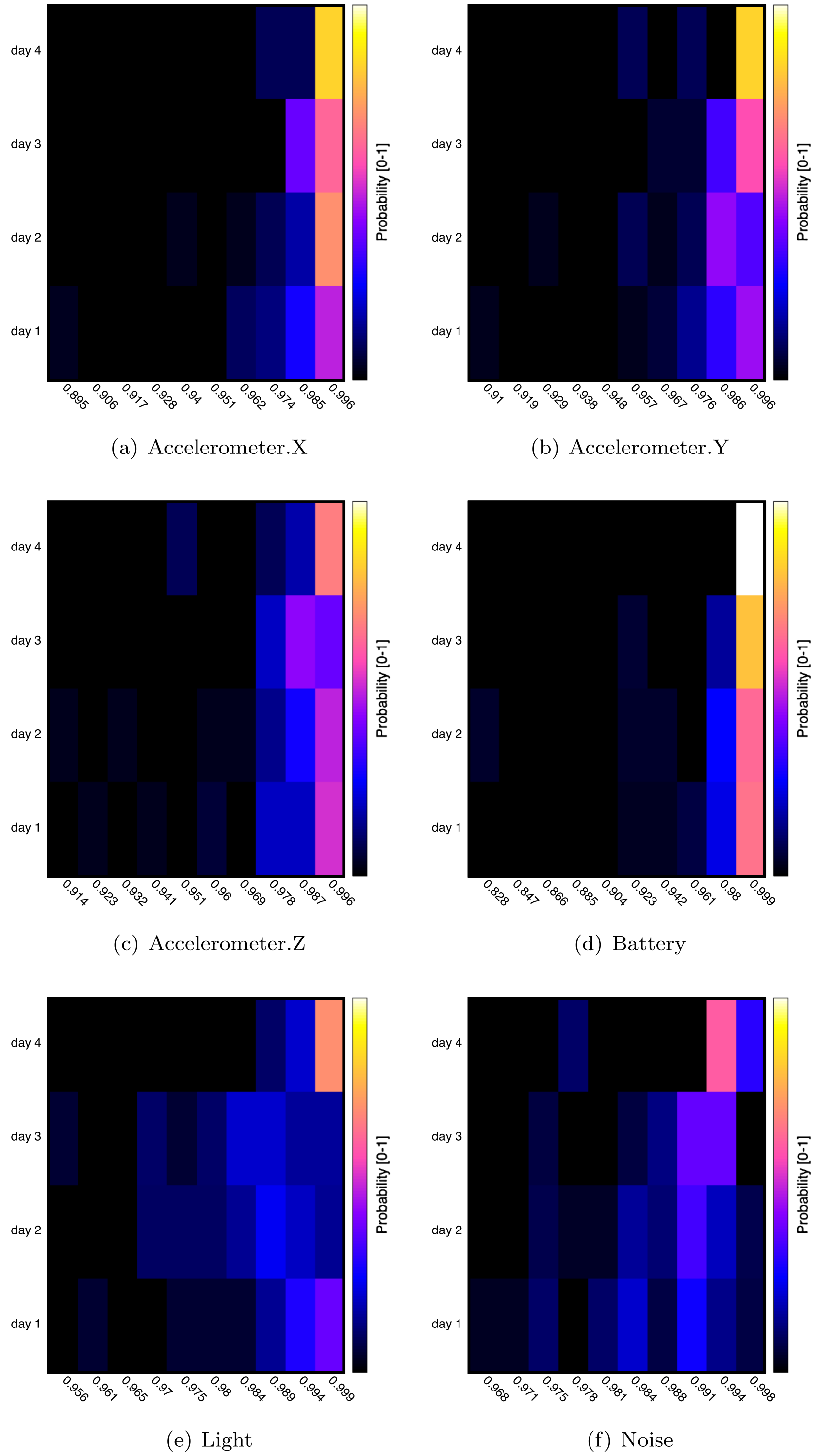
For daily summarization, the number of clusters detected varies from 1 to 10 with 2 clusters having the highest probability of 0.52. For weekly summarization, the number of clusters detected varies from 1 to 18, with 4 clusters having the highest probability of 0.3. Similarly, the probability density function of summarization for the Nervousnet project is shown in Figure 7.

The total budget of rewards for the ECBT project is derived by assuming that power utility companies return back to the consumer on average 10% of the average electricity bill, as reward for information sharing. The average electricity bill in Ireland is around €79.5 per month in 2013,[i] meaning €2.65 per day or €19.88 per week. The 10% of daily and weekly rewards corresponds to €0.265 and €1.988 per user. The total daily and weekly budget of rewards is then $0.265 * 6435 =$€1705.3 and $1.988 * 6435 =$€12792.8.

Two probability density functions $P_{\mathrm{r}}(\alpha_{i,t})$ are evaluated. The first one is the following linear probability density function:

$$P_{\mathrm{r}}(\alpha_{i,e}) = a * \alpha_{i,e} + b, \tag{10}$$

where $a = -0.36201$ and $b = 0.3803377$. The second one is an observational density function derived from the time-series data[j] of real-time tariffs. By mapping prices to summarization levels reversed proportionally and scaling up/down the average daily and weekly rewards per user according to the probability of electricity prices, the $P_{\mathrm{r}}(\alpha_{i,t})$ is approxi-



(a) Accelerometer.X

(b) Accelerometer.Y

(c) Accelerometer.Z

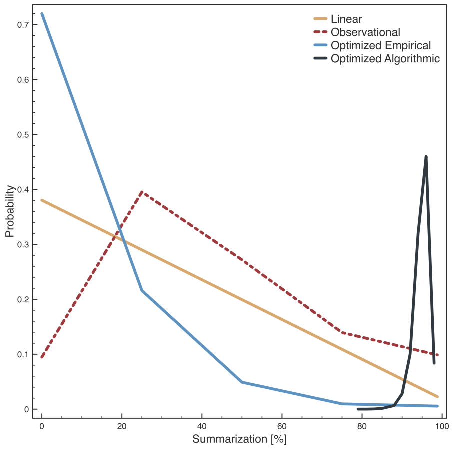
(d) Battery

(e) Light

(f) Noise

Figure 7 Probability density functions of summarization by automatically detecting the number of clusters in Nervousnet.

Figure 8 Linear, observational and optimized distributions of citizen's rewards for different summarization levels.



mated as follows:

$$P_{\mathrm{r}}(\alpha_{i,e}) = a + b * \alpha_{i,e} + c * \alpha_{i,e}^2 + d * \alpha_{i,e}^3 + e * \alpha_{i,e}^4, \qquad (11)$$

where $a = 0.09474206$, $b = 0.02920637$, $c = -9.035{,}067 \times 10^{-4}$, $d = 9.492{,}094 \times 10^{-6}$ and $e = -3.375{,}682 \times 10^{-8}$. Figure 8 illustrates the probability density functions $P_{\mathrm{r}}(\alpha_{i,t})$ used in this paper.

The rewards of a citizen i for different summarization levels can be derived from Equation (8). Figure 8 shows the linear and observational probability density functions, together with two optimized ones illustrated in Section 4.3 to make the distribution of rewards fairer.

4 Experimental evaluation

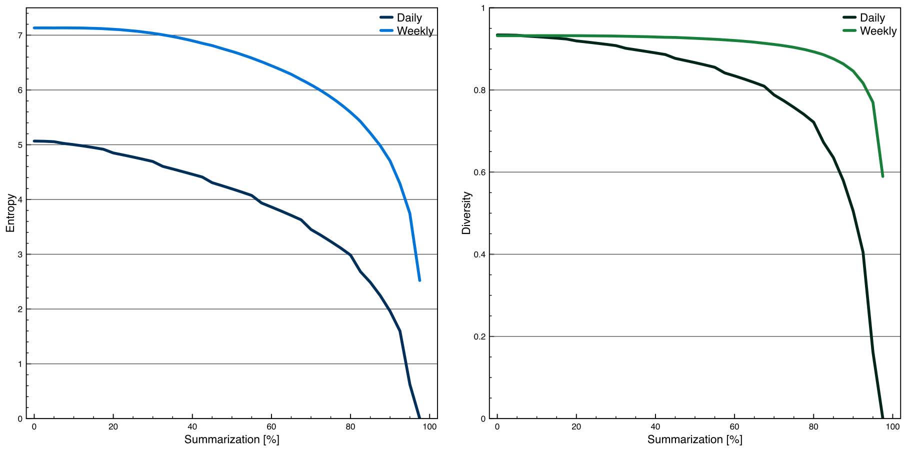
This section illustrates the experimental results for privacy, accuracy and rewards.

4.1 Privacy

Figure 9 illustrates privacy-preservation under fixed summarization levels in the ECBT project. The results concern 100% of the citizens. As summarization increases from 0%, that is the raw data, to 80%, the entropy in Figure 9(a) decreases 41.08% and 21.69% for daily and weekly summarization. An additional 10% of summarization results in an 34.34% and 15.78% drop of entropy. In Figure 9(b), the diversity decreases 22.75% and 4.22% respectively when summarization increases from 0% to 80%. The decrease is 29.96% and 5.23% respectively for an additional 10% summarization.

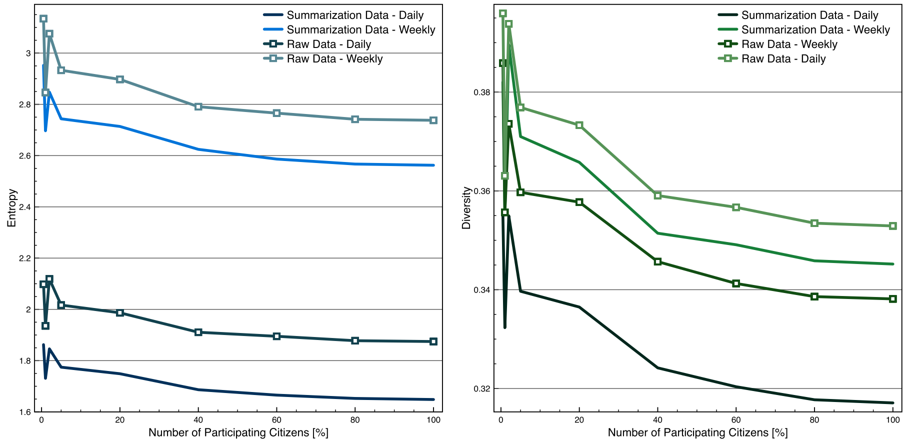
Figure 10 illustrates privacy-preservation under empirical summarization levels in the ECBT project. The results concern a varying number of participating citizens. This figure confirms the findings of Figure 9 and it additionally shows how privacy-preservation is affected when an increasing number of citizens participate in social sensing. As participation increases from 0.5% to 100%, entropy decreases 11.29% for daily and 13.22% for weekly summarization as shown in Figure 10(a). Similarly in Figure 10(b) diversity decreases 11.27% and 10.62% respectively.

Figure 11 illustrates privacy-preservation under algorithmic summarization levels in the ECBT project. The results concern a varying number of participating citizens. Similar



(a) Entropy **(b)** Diversity

Figure 9 Privacy-preservation under fixed summarization levels in the ECBT project.

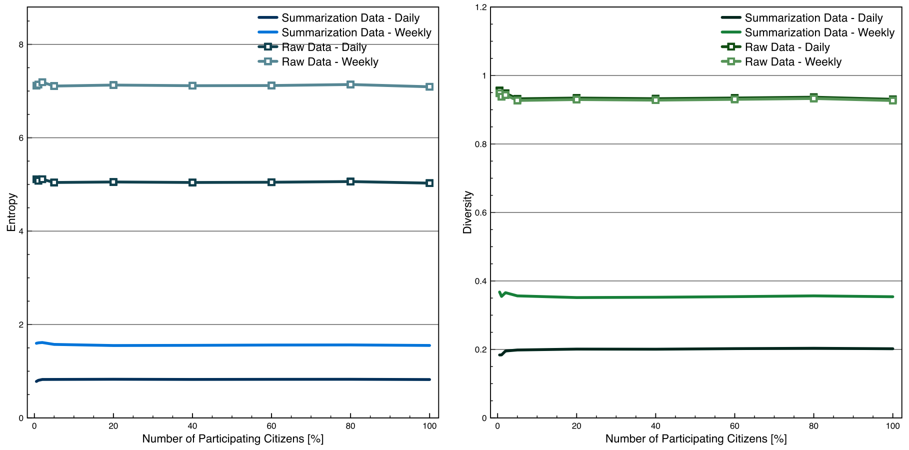


(a) Entropy **(b)** Diversity

Figure 10 Privacy-preservation under empirical summarization levels in the ECBT project.

to the empirical selections of summarization levels, the algorithmic selections improve privacy-preservation, however, the entropy and diversity for daily summarization with 100% of participating citizens are 50.18%, 36.28% lower than the ones in empirical summarization. Respectively, for weekly summarization they are 39.5% lower and 4.7% higher. This difference is because of the lower number of clusters in the algorithmic summarization.

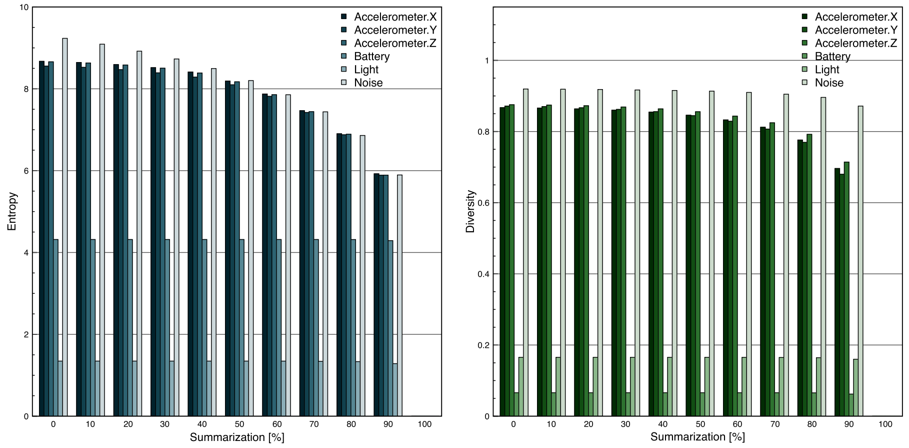
Figure 12 illustrates privacy-preservation under fixed summarization levels in the Nervousnet project. The results concern 100% of the citizens. A summarization of 0% corresponds to the raw data. Entropy and diversity decrease for all sensors except the light and battery sensors whose values do not significantly vary. The entropy of the accelerometer and noise sensor decrease 20.16%, 25.7% respectively as summarization increases from 0% to 80%. The respective decrease for diversity is 10.56% and 2.56%.



(a) Entropy

(b) Diversity

Figure 11 Privacy-preservation under algorithmic summarization levels in the ECBT project.

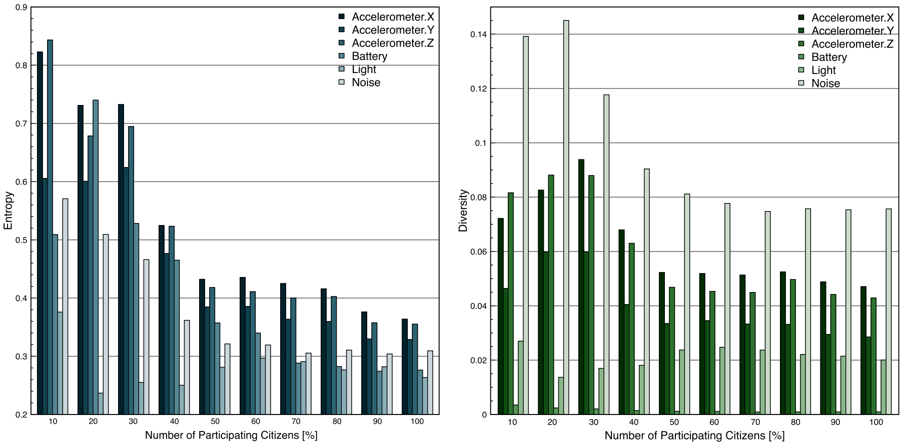


(a) Entropy

(b) Diversity

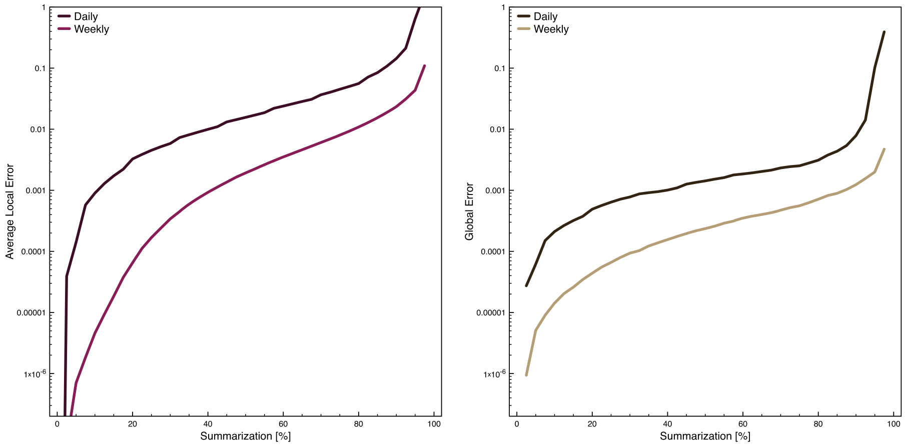
Figure 12 Privacy-preservation under fixed summarization levels in the Nervousnet project.

Figure 13 illustrates privacy-preservation under algorithmic summarization levels in Nervousnet. The results concern a varying number of participating citizens. The entropy shows a similar decreasing trend to the empirical summarization levels of the ECBT project. However, the actual values of each sensor vary significantly. Accelerometer shows the highest entropy of 0.35, whereas light sensor the lowest one of 0.26. In contrast, noise sensor has the highest diversity of 0.076 and battery sensor the lowest ones of 0.001. The accelerometer and noise sensors result in richer informational content given that changes in social activity are likely to influence the sensors values significantly. In contrast, battery and light sensors are less likely to generate rich informational content given that social activity may not trigger new sensor data. For example, a smart phone in the pocket of a walking citizen generates rich accelerometer data, but close to zero light values.



(a) Entropy **(b)** Diversity

Figure 13 Privacy-preservation under algorithmic summarization levels in the Nervousnet project.



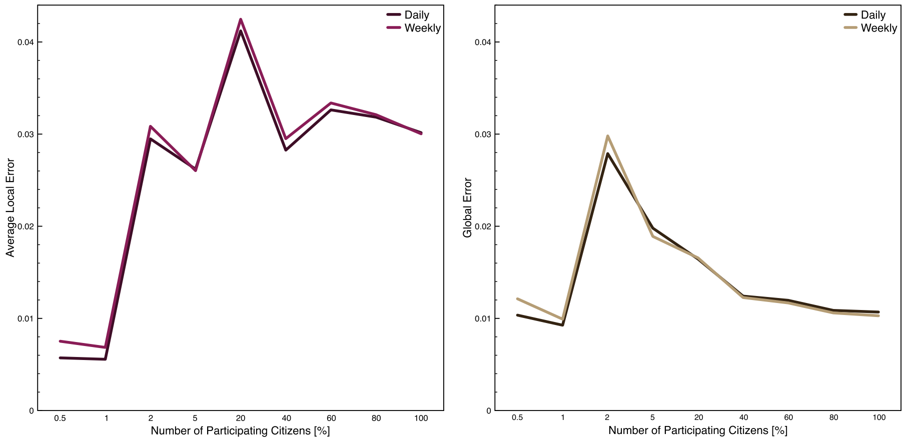
(a) Average Local Error **(b)** Global Error

Figure 14 Accuracy under fixed summarization levels in the ECBT project.

These variations among metrics confirm that privacy is a multi-dimensional concept and one metric cannot adequately quantify the dynamics of data and the perception of citizens on them.

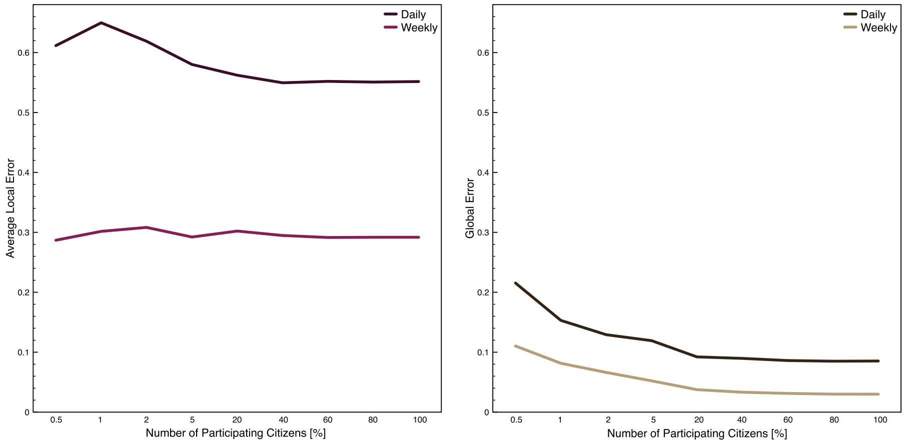
4.2 Accuracy

Figure 14 illustrates the accuracy of summation under fixed summarization levels in the ECBT project. 100% of the citizens are counted. The global error for daily summarization is 83.44% lower on average than the average local errors. This indicates an accurate summation regardless of the summarization performed. Inaccuracies only appear at very high levels of summarization. The daily average local error is 0.06% for 80% summarization, whereas, it becomes 0.14% for 90% summarization. For the summarization values, the daily global error is 0.003% and 0.008% respectively. The striking difference between



(a) Average Local Error **(b)** Global Error

Figure 15 Accuracy under empirical summarization levels in the ECBT project.



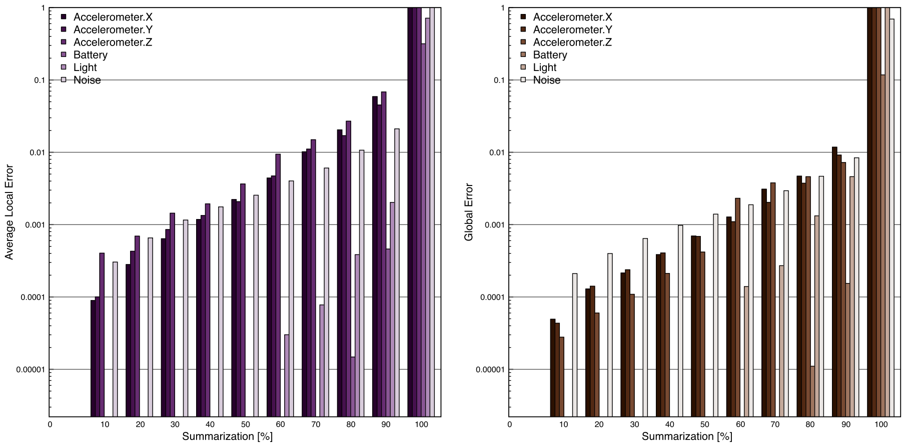
(a) Average Local Error **(b)** Global Error

Figure 16 Accuracy under algorithmic summarization levels in the ECBT project.

the average local and global errors occurs because of the cancellations in the local errors occurring in aggregation.

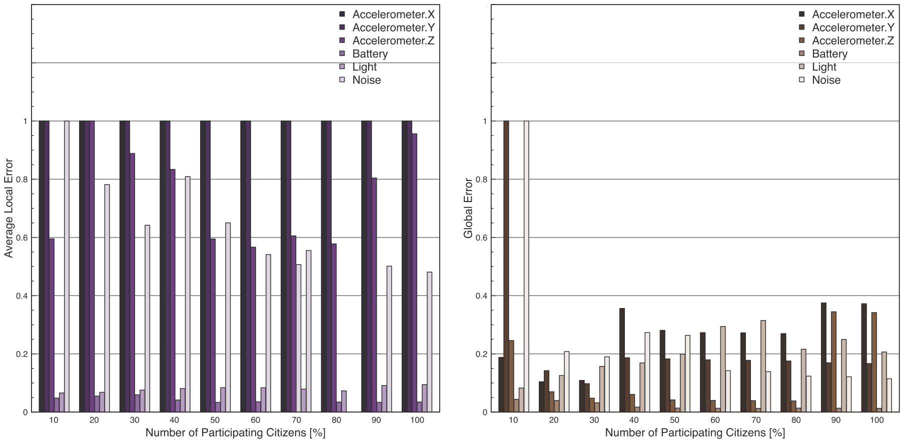
Figure 15 illustrates the accuracy of summation under empirical summarization levels in the ECBT project. The results concern a varying number of participating citizens. The average local error under 100% of participating citizens for daily and weekly summarization is 0.03 whereas the global error is 0.01 respectively.

Figure 16 illustrates the accuracy of summation under algorithmic summarization levels in the ECBT project. The results concern a varying number of participating citizens. Results confirm the trend of Figure 15, however, the average local error and global error for daily summarization and 100% of the participating citizens are 94.55% and 87.41% higher than these of the empirical summarization. Respectively for weekly summarization they



(a) Average Local Error **(b)** Global Error

Figure 17 Accuracy under fixed summarization levels in the Nervousnet project.



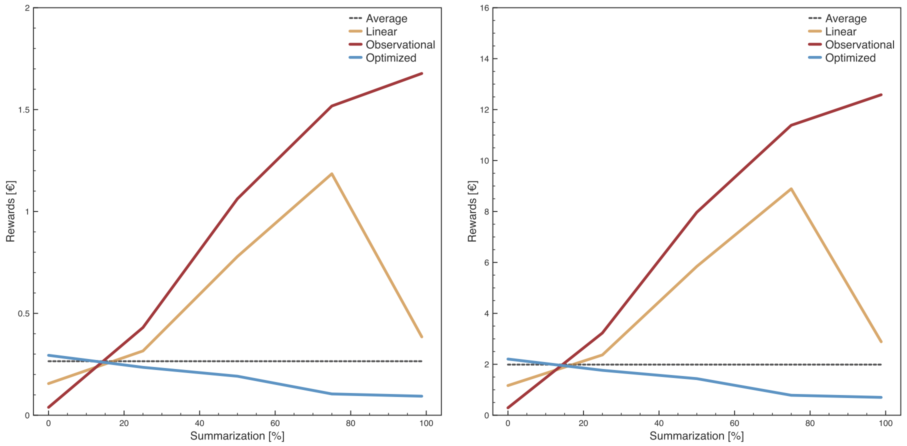
(a) Average Local Error **(b)** Global Error

Figure 18 Accuracy under algorithmic summarization levels in the ECBT project.

are 89.71% and 65.67%. The overall lower number of clusters in algorithmic summarization explains this difference.

Figure 17 shows the accuracy of average under fixed summarization levels in Nervousnet. 100% of the citizens are counted. Both errors show an approximate linear increase in the logarithmic scale as summarization increases. The global error is on average 76.81%, 47.57% and 58.12% lower than the average local error for the accelerometer, battery and noise. The significant number of zero values in the light sensor does not allow a meaningful estimate. Compared to the ECBT dataset, the error cancellations occur at a lower scale in the Nervousnet dataset that is an expected observation given the lower number of participating citizens in the latter.

Figure 18 illustrates the accuracy of average under algorithmic summarization levels in the Nervousnet project. The results concern a varying number of participating citizens.



(a) Daily

(b) Weekly

Figure 19 Rewards of a citizen under empirical summarization levels in the ECBT project.

The global error decreases 52.57% on average for all sensors in the range 10%-100% of participating citizens. The average local error decreases respectively 3.86%.

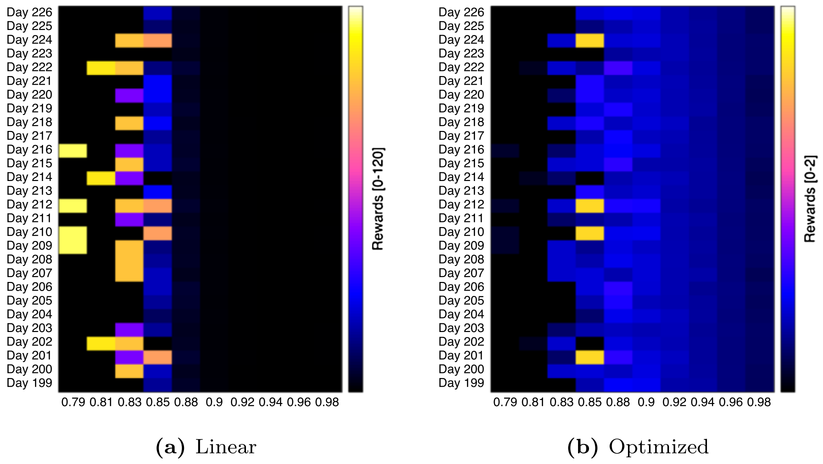
4.3 Rewards

Figure 19 illustrates the daily and weekly rewards that a citizen receives given empirical summarization levels, as defined by Equation (8). The objective of the self-regulatory information sharing is to distribute rewards fairly given the summarization level selected by each citizen. A fair distribution refers to high rewards to low summarization and low rewards to high summarization.

When $P_s() = P_r()$, rewards are distributed independent of the summarization level selected. This corresponds to the average of €0.265 and €1.988 for daily and weekly rewards. This scenario corresponds to an ideal egalitarian participation in which citizens act altruistic and rewards are not an incentive for them to increase or decrease their summarization level. The linear and observational probability density functions of Figure 8 result in a very counter-intuitive distribution of rewards in respect to fairness as shown in Figure 19. This is because of the very high number of citizens who choose a low summarization level as shown in Figure 4.

This paper contributes two optimized probability density functions, illustrated in Figure 8, one for empirical and one for algorithmic summarization levels. These functions are constructed for data providers to make the distribution of rewards fairer and incentivize low summarization levels with higher rewards. The optimization process starts by setting $P_s() = P_r()$ and redistributing probability mass from higher summarization levels to the lowest ones in $P_r()$ of Equation (8).[k] The process repeats until the distribution of rewards approximates a linear decrease as shown in Figure 19. In the resulting distribution, citizens with 0% summarization level receive 75.17% higher rewards than citizens with 100% summarization.

Figure 20 shows the rewards a citizen receives under algorithmic summarization levels in the ECBT project. In this case, algorithmic summarization levels vary over time. Under the linear reward distribution of Figure 8, rewards are distributed highly unequally



(a) Linear (b) Optimized

Figure 20 Rewards of a citizen under algorithmic summarization levels in the ECBT project.

Table 3 A comparison of the proposed system with related work in the light of the six future research challenges to be tackled as illustrated in earlier survey papers [9, 10]

Related work	Participants in privacy equation	Composable privacy solutions	Privacy, performance & data fidelity trade-offs	Measurable privacy	Standards for privacy research	Holistic architecture blueprints
[23]	✓			✓		
[18]	✓	✓		✓		
[17]		✓	✓	✓		
[24]				✓	✓	
[25]		✓			✓	
[26]					✓	
[27]		✓	✓	✓	✓	✓
[28]	✓		✓	✓		
[29]	✓		✓	✓	✓	
[30]				✓	✓	
[31]	✓			✓	✓	
[32]	✓			✓		
Proposed	✓	✓	✓	✓	✓	✓

with only 59.95% of the total budget spent as illustrated in Figure 20(a). Citizens with the lower summarization values receive rewards magnitudes higher than citizens with high summarization values. In contrast, the optimized algorithmic $P_r()$ of Figure 8 incentivizes citizens more fairly as shown in Figure 20(b).

5 Comparison with related work

An extensive review of privacy threats and possible countermeasures in participatory sensing applications is earlier illustrated [9, 10]. The authors conclude on six research challenges for future work to tackle: (i) including the participants in the privacy equation, (ii) providing composable privacy solutions, (iii) trade-offs between privacy, performance and data fidelity, (iv) making privacy measurable, (v) defining standards for privacy research and (vi) holistic architecture blueprints. Compared to related work, the proposed self-regulatory information sharing system contributes to all of these challenges as outlined in Table 3.

This work empirically shows how the preferences of citizens can be mapped to measurable and composable privacy protection levels. These preferences can be self-determined dynamically and change over time. Moreover, the proposed self-regulatory mechanism is generic as it does not depend on specific sensor data or application. Trade-offs between privacy, accuracy of computations and costs are quantified under different regulatory settings. Quantitative results are illustrated with metrics that can be standardized for different types of privacy research. This work sets the foundations for a holistic architecture blueprint given that self-regulation of information sharing is modeled as a supply-demand system. This approach opens up new research opportunities stemmed from scientific fields such as machine learning, computational markets and evolutionary game-theory. The rest of this section elaborates on related work and draws comparisons with the proposed system for self-regulatory information sharing.

Two different privacy concepts, k-anonymity and l-diversity, are earlier investigated, showing how privacy models can be applied to protect users' spatial and temporal privacy in the context of participatory sensing [30]. Similarly, PoolView [31] relies on data perturbation on the client-side to ensure individuals' privacy in aggregation. The accuracy of the analytics is to certain extent fixed and bound to the adopted distribution functions for perturbing the data. In contrast to these methods, the proposed local summarization technique can be configured to provide different levels of privacy and rewards among citizens for a broad range of sensor types as shown in this paper.

Privacy in participatory sensing is earlier discussed [32] by studying how reputation values, representing a level of trust on the contributed data, are transferred between anonymous contributors. Furthermore, the authors also introduce a centralized reputation anonymization scheme, which aims at preventing leakage of privacy due to the reputation information used. On the contrary, the summarization technique introduced in this paper is performed locally and data trustworthiness is regulated using rewards.

MyExperience [23] identifies both objective and subjective data from mobile computing activities. More specifically, MyExperience combines a variety of techniques, including passive logging of device usage, user and environmental sensor readings, subjective user feedback, as well as context-triggered user experience. However, MyExperience does not provide summarization techniques in this context. In contrast, this research introduces a non-proprietary system that can collect a wide range of privacy-preserving sensor data on individuals' personal mobile devices.

An investigation on a user-cantered framework is carried out, with particular emphasis on the support of individual users with privacy awareness and control in ubiquitous computing environments [18]. In particular, the author investigates physical privacy aspects in terms of territorial privacy. The main discussion focuses on the following: awareness of privacy implications, privacy concerns and human factors and territorial privacy models. The author subsequently proposes a model instantiation and the concept of channel policies as the instantiation of the territorial privacy model. A user interface for user-centered privacy management is discussed, which is based on an iterative process designed by several user studies and online surveys. An important remark is that privacy rules and regulations should be mapped onto concrete regulatory privacy systems such as the one proposed in this paper. Furthermore, the design and development of a simplified user interface is required for non-technical users.

A recommender system implemented as a mobile application based on modern portfolio theory is earlier introduced with specific security and privacy awareness. It automatically assesses the security risk of mobile applications [17]. In fact, a mobile application can pose the security risk due to insecure access permissions. Furthermore, the recommendations are based on both the popularity of the application and the users' security preferences. The experiments on a large-scale real-world dataset from Google Play validate the effectiveness and efficiency of the proposed recommendation framework. On the contrary, this paper focuses on summarization that is autonomously regulated by the user. The results of this work could be used to build a privacy recommender systems that personalizes the summarization values for each citizen.

Privacy issues are one of the major concerns among the majority of users, especially regarding the way corporations collect personal data. In particular, there is a widespread consensus on the fact that privacy policies are too complex and ambiguous to fully understand. Earlier research determines that users purchasing both non-privacy-sensitive and privacy-sensitive items are willing to pay a premium for privacy when privacy information is more comprehensive and relevant [24].

A more effective analysis of policies and an assessment of ambiguity in privacy policies is earlier introduced by employing informational lexicon from manual, human annotations and an entity extractor based on part-of-speech tagging [25]. The authors measure the terminological reuse across a variety of policies. The lexicon reached a saturation limit of between 31-78% in three domains, suggesting an incomplete lexicon. However, the authors argue that the lexicon can still improve textual analysis of privacy policies by assessing common words and text fragments. Similarly, another approach introduces a semi-automatic extraction of privacy features from natural language privacy polices [26]. These features are presented to users in a comprehensive and friendly format. This facilitates more informed privacy decision-making during the users' interaction with a variety of websites. Trends in the content of web privacy policies are systematically identified.

In another work [27], multi-party data flow requirements are modeled using descriptive logic. Conflicts and violations of the privacy principles are identified, as well as two patterns for balancing privacy and data use in specification requirements. Furthermore, the authors' analysis of automation reasoning over models of descriptive logic demonstrates that reasoning over complex compositions of multi-party systems exhibit efficiency and scalability. The authors carry out an evaluation on an empirical case-study by examining the data practices of the Waze mobile application, Facebook Login, Amazon Web Services, and Flurry.com.

Furthermore, the privacy settings of mobile applications cannot assess people's perceptions of whether a given action is legitimate, or how that action makes them feel with respect to privacy. A model for privacy as expectations is introduced [28] that uses crowdsourcing to capture users' expectations of which sensitive information mobile applications utilize. Results show that users feel more comfortable when they are informed about the reasons sensitive information is required. Uncertainties negatively affect users regarding their privacy.

Location-sharing applications, in which users specify the conditions under which they are willing to allow other users to detect their locations are earlier studied [29]. The authors define canonical policies to describe user-specific elements, as well as canonical places, based on decision-tree and clustering algorithms. The results suggest that a more targeted choice of the default canonical policies can potentially facilitate the customization of privacy settings.

6 Conclusion and future work

This paper concludes that between the one extreme of intruding privacy to collective massive scales of data and the other extreme of limiting participation in social sensing applications to protect privacy there is a trajectory of viable and sustainable solutions for self-regulatory information sharing. This paper shapes this trajectory by modeling information sharing between citizens and data aggregators as a supply-demand system supported by computational markets. The system design captures both (i) citizens' selections and (ii) aggregators' incentives about how rewards are split to different summarization levels. Results show that incentivization can be optimized to be fair. The performed experiments show that the information loss by the local summarization is higher than the information loss in the computed analytics concluding that analytics can tolerate the information loss that summarization causes. Results also quantify the influence of different sensor types in performance, confirming that privacy is a multi-dimensional concept. Therefore, self-regulatory information sharing should be highly contextualized and tailored to different data and applications. For this, the findings of this work can be used as a guide in future work.

However, this work has also limitations and open issues to address in future work. Performing a real-world social experiment to acquire more realistic citizens' preferences shall further validate this work. Every market mechanism requires effective institutions and policies to guarantee compliance, fairness and social justice. A market design should also capture social, ethical and cultural norms. Lessons learnt from other market mechanisms in finance and energy can be applicable in this new application domain [16, 33–35]. When computational markets are not a viable approach, the findings of this work are also applicable for privacy personalization via recommender systems as outlined in related work [17]. Moreover, instead of summarization based on clustering, the robustness to inference shall be studied in techniques such as the perturbations of PoolView [31], synopsis diffusion [36] and text summarization techniques [37]. Finally, citizens can contribute their computational resources to acquire the role of a data aggregator in order to participate in a fully decentralized data analytics process [38–40]. This potential is the endeavor of an alternative participatory, truly decentralized Big Data paradigm supporting digital democracy [41].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Authors contributed with the order they appear. All authors read and approved the final manuscript.

Author details
[1]Professorship of Computational Social Science, ETH Zurich, Clausiusstrasse 50, Zurich, 8092, Switzerland. [2]Department of Computing and Mathematics, University of Derby, Kedleston Road site, Derby Campus, Derby, DE22 1GB, UK. [3]Department of Computing, Edge Hill University, St Helens Road, Ormskirk, Lancashire, L39 4QP, UK.

Acknowledgements
This work is supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem' (http://www.sobigdata.eu) and the European Community's H2020 Program under the scheme 'ICT-10-2015 RIA', grant agreement #688364 'ASSET: Instant Gratification for Collective Awareness and Sustainable Consumerism'. The authors would like to thank Jesús Rodríguez-Molina for his enthusiasm in supporting this research.

Endnotes

[a] Available at https://storage.googleapis.com/think-emea/docs/article/Mobile_App_UX_Principles.pdf (last accessed: February 2016)

[b] This range can be constrained in case stricter privacy policies are required in some critical data, *e.g.*, entertainment vs. health applications.

[c] Available at http://www.ucd.ie/issda/data/commissionforenergyregulationcer/ (last accessed November 2015)

[d] For simplicity, the absolute values of the $\langle x, y, z\rangle$ accelerometer records are used.

[e] The average aggregation function is more meaningful to compute for the Nervousnet sensors, *e.g.* summed acceleration vs. average acceleration.

[f] The approximation quality is measured as follows: $R^2 = 0.9987$, $aR^2 = 0.9967$, $P = 0.001958$, $SE = 0.01366$ and $F = 510$.

[g] Two other questions are relevant here as well: (i)*My organization is interested in changing the way we use electricity if it reduces the electricity bill* and (ii) *My organization is interested in changing the way we use electricity if it helps the environment*. All three questions give similar probability distributions of answers. For simplicity, only one of them is used in the experimental evaluation.

[h] The approximation quality is measured as follows: $R^2 = 0.9945$, $aR^2 = 0.9862$, $P = 0.008298$, $SE = 0.0241$ and $F = 119.7$.

[i] Available at http://www.irishexaminer.com/ireland/electricity-and-gas-costs-up-200-in-past-year-233268.html (last accessed: November 2015)

[j] Five-minute price signals are used from the Pacific Northwest Smart Grid Demonstration project. The signals concern transmission zone '12', site '0' for the period 30.07.2013-06.08.2013.

[k] Given that $P_s()$ differs in the algorithmic summarization daily or weekly, the average probability density function is computed across time and it is the one applied in the optimization process for computing the optimized algorithmic $P_r()$.

Received: 16 November 2015 Accepted: 21 March 2016 Published online: 01 April 2016

References

1. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19(2):171-209
2. Pournaras E, Vasirani M, Kooij RE, Aberer K (2014) Decentralized planning of energy demand for the management of robustness and discomfort. IEEE Trans Ind Inform 10(4):2280-2289
3. Pournaras E (2013) Multi-level reconfigurable self-organization in overlay services. PhD thesis, Delft University of Technology
4. Hajian S, Domingo-Ferrer J, Monreale A, Pedreschi D, Giannotti F (2015) Discrimination- and privacy-aware patterns. Data Min Knowl Discov 29(6):1733-1782
5. Pedreschi D, Ruggieri S, Turini F (2013) The discovery of discrimination. In: Discrimination and privacy in the information society. Springer, pp 91-108
6. Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '08), pp 560-568
7. Getoor L, Machanavajjhala A (2013) Entity resolution for big data. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '13). ACM, New York, p 1527
8. Liu X, Lu M, Ooi BC, Shen Y, Wu S, Zhang M (2012) CDAS: a crowdsourcing data analytics system. Proc VLDB Endow 5(10):1040-1051
9. Christin D (2015) Privacy in mobile participatory sensing: current trends and future challenges. J Syst Softw. doi:10.1016/j.jss.2015.03.067
10. Christin D, Reinhardt A, Kanhere SS, Hollick M (2011) A survey on privacy in mobile participatory sensing applications. J Syst Softw 84(11):1928-1946
11. Christin D, Buchner C, Leibecke N (2013) What's the value of your privacy? Exploring factors that influence privacy-sensitive contributions to participatory sensing applications. In: 2013 IEEE 38th conference on local computer networks workshops (LCN workshops), pp 918-923
12. Pournaras E, Vasirani M, Kooij RE, Aberer K (2014) Measuring and controlling unfairness in decentralized planning of energy demand. In: 2014 IEEE international energy conference (ENERGYCON), pp 1255-1262
13. Helbing D, Hennecke A, Shvetsov V, Treiber M (2002) Micro- and macro-simulation of freeway traffic. Math Comput Model 35(5-6):517-547
14. Sun H, Florio VD, Gui N, Blondia C (2009) Promises and challenges of ambient assisted living systems. In: 2009. ITNG '09. Sixth international conference on information technology: new generations. IEEE, pp 1201-1207
15. Jaynes ET (1982) On the rationale of maximum-entropy methods. Proc IEEE 70(9):939-952
16. Roth AE (2002) The economist as engineer: game theory, experimentation, and computation as tools for design economics. Econometrica 70:1341-1378
17. Zhu H, Xiong H, Ge Y, Chen E (2014) Mobile app recommendations with security and privacy awareness. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 951-960
18. Könings B (2015) User-centered awareness and control of privacy in Ubiquitous Computing. PhD thesis, Universität Ulm, Ulm
19. Christin D, Engelmann F, Hollick M (2014) Usable privacy for mobile sensing applications. In: Information security theory and practice. Securing the Internet of things. Lecture notes in computer science, vol 8501. Springer, Berlin, pp 92-107
20. Christin D, Reinhardt A, Hollick M, Trumpold K (2012) Exploring user preferences for privacy interfaces in mobile sensing applications. In: Proceedings of the 11th international conference on mobile and ubiquitous multimedia (MUM '12). ACM, New York, pp 14:1-14:10

21. Pournaras E, Moise I, Helbing D (2015) Privacy-preserving ubiquitous social mining via modular and compositional virtual sensors. In: 2015 IEEE 29th international conference on advanced information networking and applications (AINA), pp 332-338
22. Jain AK, Maheswari S (2012) Survey of recent clustering techniques in data mining. Int J Comput Sci Manag Res 3:72-78
23. Froehlich J, Chen MY, Consolvo S, Harrison B, Landay JA (2007) MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the 5th international conference on mobile systems, applications and services. ACM, pp 57-70
24. Tsai JY, Egelman S, Cranor L, Acquisti A (2011) The effect of online privacy information on purchasing behavior: an experimental study. Inf Syst Res 22(2):254-268
25. Sadeh N, Acquisti A, Breaux TD, Cranor LF, McDonalda AM, Reidenbergb JR, Smith NA, Liu F, Russellb C, Schaub F, Wilson S (2013) The usable privacy policy project: combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Technical report CMU-ISR-13-119, Carnegie Mellon University
26. Breaux TD, Smullen D, Hibshi H (2015) Detecting repurposing and over-collection in multi-party privacy requirements specifications. In: 2015 IEEE 23rd international requirements engineering conference (RE). IEEE, New York, pp 166-175
27. Lin J, Amini S, Hong JI, Sadeh N, Lindqvist J, Zhang J (2012) Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, New York, pp 501-510
28. Ravichandran R, Benisch M, Kelley PG, Sadeh NM (2009) Capturing social networking privacy preferences. In: Privacy enhancing technologies. Springer, pp 1-18
29. Bhatia J, Breaux TD (2015) Towards an information type lexicon for privacy policies. In: 2015 IEEE eighth international workshop on requirements engineering and law (RELAW). IEEE, New York, pp 19-24
30. Huang KL, Kanhere SS, Hu W (2010) Preserving privacy in participatory sensing systems. Comput Commun 33(11):1266-1280
31. Ganti RK, Pham N, Tsai Y, Abdelzaher TF (2008) PoolView: stream privacy for grassroots participatory sensing. In: Proceedings of the 6th ACM conference on embedded network sensor systems (SenSys '08). ACM, New York, pp 281-294
32. Huang KL, Kanhere SS, Hu W (2012) A privacy-preserving reputation system for participatory sensing. In: Proceedings of the 2012 IEEE 37th conference on local computer networks (LCN 2012). IEEE Comput. Soc., Washington, pp 10-18
33. Biais B, Glosten L, Spatt C (2005) Market microstructure: a survey of microfoundations, empirical results, and policy implications. J Financ Mark 8(2):217-264
34. Chan LMA, Shen ZJ, Simchi-Levi D, Swann J (2004) Coordination of pricing and inventory decisions: a survey and classification. In: Simchi-Levi D, Wu SD, Shen Z (eds) Handbook of quantitative supply chain analysis. International series in operations research and management science, vol 74. Springer, Berlin, pp 335-392
35. David AK, Wen F (2000) Strategic bidding in competitive electricity markets: a literature survey. In: 2000 IEEE power engineering society summer meeting, vol 4, pp 2168-2173
36. Roy S, Conti M, Setia S, Jajodia S (2012) Secure data aggregation in wireless sensor networks. IEEE Trans Inf Forensics Secur 7(3):1040-1052
37. Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Mining text data. Springer, Berlin, pp 43-76
38. Pournaras E, Warnier M, Brazier FM (2013) A generic and adaptive aggregation service for large-scale decentralized networks. Complex Adapt Syst Model 1(1):19
39. Boutsis I, Kalogeraki V (2013) Privacy preservation for participatory sensing data. In: Proceedings of the 2013 IEEE international conference on pervasive computing and communications (PerCom). IEEE, New York, pp 103-113
40. Dürr M, Wiesner K (2011) A privacy-preserving social P2P infrastructure for people-centric sensing. In: Luttenberger N, Peters H (eds) 17th GI/ITG conference on communication in distributed systems (KiVS 2011). OpenAccess series in informatics (OASIcs), vol 17. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl, pp 176-181
41. Helbing D, Pournaras E (2015) Build digital democracy. Nature 527:33-34