View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

Consent-driven data use in crowdsensing platforms: When data reuse meets privacy-preservation

Mariem Brahem, Guillaume Scerri, Nicolas Anciaux, Valerie Issarny

To cite this version:

Mariem Brahem, Guillaume Scerri, Nicolas Anciaux, Valerie Issarny. Consent-driven data use in crowdsensing platforms: When data reuse meets privacy-preservation. PerCom 2021 - IEEE International Conference on Pervasive Computing and Communications, Mar 2021, Kassel / Virtual, Germany. hal-03097047v3

HAL Id: hal-03097047

https://hal.inria.fr/hal-03097047v3

Submitted on 11 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Consent-driven data use in crowdsensing platforms: When data reuse meets privacy-preservation

Mariem Brahem[1,2], Guillaume Scerri[1,2], Nicolas Anciaux[1,2], and Valerie Issarny[3]

[1]Petrus team, Inria, France
[3]Mimove team, Inria, France
[2]University of Versailles Saint-Quentin-en-Yvelines, France
{mariem.habibi, guillaume.scerri, nicolas.anciaux, valerie.issarny}@inria.fr

***Abstract*—Crowdsensing is an essential element of the IoT; it allows gathering massive data across time and space to feed our environmental knowledge, and to link such knowledge to user behavior. However, there are major obstacles to crowdsensing, including the preservation of privacy. The consideration of privacy in crowdsensing systems has led to two main approaches, sometimes combined, which are, respectively, to trade privacy for rewards, and to take advantage of privacy-enhancing technologies "anonymizing" the collected data. Although relevant, we claim that these approaches do not sufficiently take into account the users' own tolerance to the use of the data provided, so that the crowdsensing system guarantees users the expected level of confidentiality as well as fosters the use of crowdsensing data for different tasks. To this end, we introduce the *ℓ-completeness* property, which ensures that the data provided can be used for all the tasks to which their owners consent as long as they are analyzed with $\ell - 1$ other sources, and that no privacy violations can occur due to the related contribution of users with less stringent privacy requirements. The challenge, therefore, is to ensure *ℓ-completeness* when analyzing the data while allowing the data to be used for as many tasks as possible and promoting the accuracy of the resulting knowledge. We address this challenge with a clustering algorithm sensitive to the data distribution, which is shown to optimize data reuse and utility using a dataset from a deployed crowdsensing application.**

***Index Terms*—Crowdsensing, Privacy, Consent.**

I. Introduction

Mobile crowdsensing is an essential element of the Internet Of Things (IoT) as it allows gathering tremendous data across time and space at low cost [1]. Indeed, thanks to the democratization of smartphones that embed increasingly rich sensing capabilities, we are able to sense a large portion of the physical environment and further relate the observed phenomena with human behavior. Various applications illustrate the benefit of mobile crowdsensing toward better informing and enhancing, e.g., environmental monitoring and awareness [2], public health monitoring and policy [3] or traffic management [4]. Still, mobile crowdsensing comes with tremendous challenges for it to be widely adopted and to effectively feed today's AI-powered systems. Challenges span: the ability to embark a sufficiently large crowd to gather the required spatio-temporal knowledge [5]; ensuring the quality of data through supporting trust management [6], context-awareness and filtering [7]; the resource-efficiency of the overall process from the data collection up to its overall aggregation [8]; and, last but not least, enforcing privacy [9].

This paper specifically focuses on the challenge of privacy-preserving mobile crowdsensing since it is the most critical one for crowdsensing to be a powerful technology that both brings valuable knowledge and serves the public good. The various dimensions of the challenge together with supporting solutions have been the focus of several surveys among which: [9], [10], [11], [12]. Relevant studies include leveraging state of the art Privacy-Enhancing Technologies (PET) to enforce related privacy metrics [13], [14]. For instance, the early work in [15] leverages decentralization together with spatial $k-$anonymity for privacy-aware task assignment. More recently, the work in [16] introduces a mechanism based on differential-privacy and distortion-privacy to guarantee that the gathering of location-based measurements is not at the expense of location privacy for the contributing users, while reducing the resulting loss of data quality. Still, leveraging PET results in the obfuscation of the crowdsensed data and thus impacts the significance of the knowledge that may be analyzed. To overcome the loss of knowledge accuracy, a significant body of research concentrates on dealing with the tension between accuracy and privacy [17]. Proposed solutions include privacy-aware auction-based approaches so that the contributors get rewarded for the loss of privacy [18], [19], [20]. In a nutshell, the more users accept to provide close-to-actual observations, the more the crowdsensing system gathers accurate knowledge and contributors get rewarded. Other approaches leverage decentralization for managing the information about the contributing users [21]. However, they focus on privacy-preservation at the time of task assignment, and do not address the complementary issue of privacy-preserving data collection.

Overall, the state of the art of privacy-preserving mobile crowdsensing provides a number of advanced protocols that may be combined toward enforcing some level of privacy. And, whatever is the crowdsensing protocol implemented, it comes with a necessary trade-offs between knowledge accuracy and privacy guarantees. The challenge is then for the crowdsensing system to get the best out of the contributed data. Part of the solution lies in the elicitation of application-specific data analyses to reduce the loss of accuracy [22]. However, we argue that it is as important to foster the re-use of data across tasks, as also advocated by the IoT data marketplace trend [23]. Indeed, this results in the enhanced resource-efficiency of mobile crowdsensing. This is known as multi-tasking in participatory sensing, for which recent studies focus on optimizing the allocation of tasks from the perspective of the task organizer [24], resp. task participant [25]. Our work distinguishes itself from, and complements, related research by concentrating on *fostering privacy-preserving data reuse* in multi-tasking, mobile crowdsensing systems. We specifically focus on participatory sensing where users explicitly register for possible participation to tasks although our approach could easily be adapted to opportunistic sensing.

In a nutshell, the research question we address is: "*How to foster the reuse of crowdsensed data across various eligible tasks while still guaranteeing the right level of privacy to the contributing users*?". A first design choice that we make is to address privacy preservation according to the user's consent to the use of their data. Indeed, digital consent is an integral part of privacy management in computing platforms [26]. We thus introduce the *ℓ-completeness* property for mobile crowdsensing platforms, which defines the extent to which mobile users consent to the reuse of the contributed observations. Precisely, the *ℓ-completeness* property enforces that: (1) the user's data are analyzed with at least $\ell - 1$ other data sources in all the tasks the data contribute to, and (2) no individual knowledge may be inferred due to the participation to many tasks that may not involve the same contributors. The proposed consent-driven property directly derives from the properties associated with sample size determination [27] in relation with privacy [28]. The challenge is then to aggregate the users' contributed data in as many allowed and relevant tasks as possible while guaranteeing *ℓ-completeness*. This leads us to introduce a supporting clustering algorithm that is sensitive to the data distribution so as that it optimizes both data reuse and utility (*aka* knowledge accuracy). The contribution of this paper is as follows:

- We define the *ℓ-completeness* property, starting with the analysis of the problem it addresses (§ II) followed by its formalization (§ III).
- We introduce a base clustering approach implementing the *ℓ-completeness* property, which we evaluate in terms of data reuse and utility, using a dataset from a deployed crowdsensing application (§ IV).
- The result of the above evaluation shows that a base implementation of *ℓ-completeness* is at the expense of data utility. We overcome the shortcoming by enhancing the clustering algorithm with the knowledge of the data distribution. The evaluation using our supporting dataset shows that the algorithm increases both the reuse and utility of the contributed data (§ V).

Finally, we position our contribution with respect to related work (§ VI) and offer conclusion (§ VII).

II. Problem Analysis

Consent is one of the legal frameworks set in many places (e.g., GDPR [29], CCPA [30]) as a precondition for any processing of personal data. In this context, *informed and specific* consent requires that users be informed of the type of data collected, the identity of the recipient of the results, and the precise nature and purpose of the task. In particular, these legal terms are intended to act as a safeguard against function creep and *data reuse*[1]. Thus, personal data collected for one function on the basis of consent cannot be extended or reused for another function without obtaining new consent. Defects in consent are considered a cause of nullity in many countries (e.g., see [32] for European countries such as France, Germany or UK).

Focusing on multi-tasking participatory sensing systems, consent-driven participation allows users to specify the tasks to which they consent to contribute with mobile observations according to their privacy requirements. Without loss of generality, we consider that a task is defined by: the function f (code) applied to the collected data of the specified type $\mathcal{S}$, the time period Δ during which each participant contributes observations, a minimal number ℓ of participants required to provide contributions to execute f[2]. The value of ℓ is deemed critical to both: (i) obtain a useful result (e.g., the evaluation of the noise level in a street requires the analysis of several contributions [33]), and (ii) protect the privacy of the participants as their individual contributions get aggregated with the ones of the $\ell - 1$ others (e.g., national agencies and data research centers impose a minimum number of individuals to be taken into account [28] when producing any aggregate –table, graph or map based on aggregate values– for research purposes: for instance, the INSEE confidentiality guide [34] and CASD rules [35] impose a minimum of 11 individuals for any computation based on tax data or 5 individuals for social data). The value of ℓ (resp. Δ) is task-dependent; hence for the sake of simplicity but without loss of generality, ℓ (resp. Δ) is aggregated as the maximum of the ℓs (resp. Δs) of all the tasks involving common contributors. Following, each task is associated with a Manifest that summarizes how the contributed data is consumed and thereby allows users to provide an informed and specific consent for the task. The Manifest is specified using a dedicated language such as, e.g., the AnonyTL language [36] introduced by the AnonySense privacy-aware system for opportunistic sensing [37]. We specifically assume the following specification for a task and its manifest:

Definition 1 (**Task and associated Manifest**). *A task T is defined as a quadruple $< f, \mathcal{S}, \Delta, \ell >$ such that f is a function (code) executed on an input set with non empty contributions*

[1] According to European Commission [31]: "*If your company/organisation has collected the data on the basis of consent (...) no further processing beyond what is covered by the original consent or the provisions of the law is possible. Further processing would require obtaining new consent*"

[2] Additional elements may characterize the data consumption by the task – e.g., the frequency of execution or the retention period of the data. Considering such parameters is area for future work.

of type $\mathcal{S}$ produced by at least ℓ consenting participants over a time period Δ. We denote by O_T the result of T. The manifest $M(T)$ is a declarative and intelligible representation of T regarding the consumption of the gathered data so that users may provide an informed and specific consent for it.

We consider a *fully trusted* multi-tasking crowdsensing system[3]. Given a set of m tasks $\mathcal{T} = \{T_i\}_{0<i\leq m}$ and associated manifests, managed and advertised by such trusted system, any user can consent to a desired subset of $\mathcal{T}$. We denote the consents of a user u with the tuple $C_u =< b_1, \ldots, b_m >$ where $b_i = 1$ if the user consents to T_i and $b_i = 0$ otherwise. By giving consents, a user accepts contributing to (only) the set of results $\{O_{T_i} = T_i.f(s_{U_i})\}_{T_i \in \mathcal{T}|C_u.b_i=1}$, with s_{U_i} being the related contributions of type $T_i.\mathcal{S}$ from a set U_i of users consenting to T_i that includes u.

Combining the contributions of users who tolerate different disclosure policies –as defined from their consents to multiple tasks– creates the risk of unintended secondary uses and may result in *defects in consent*. Consider the knowledge $\mathcal{K}$ that the participatory sensing system may disclose from the set of tasks $\mathcal{T}$, provided the table $\mathcal{C} = \{< u, C_u >\}_{u\in\mathcal{U}}$ of consents of the set of registered users $\mathcal{U}$ to the tasks (e.g., see Table I), i.e., $\mathcal{K} = \{O_{T_i}\}_{T_i\in\mathcal{T}|\mathcal{C}.b_i}$. The system must guarantee that it conforms to $\mathcal{C}$ while computing and delivering $\mathcal{K}$. However, we claim that meeting this constraint only is not sufficient to guarantee that there is no defect in consent. For example, consider the consents of 10 users to 3 tasks represented in Table I. All 10 users consent to T_1 (e.g., a task computing the average noise at a given location to request the city government to take appropriate measures to reduce the nuisance). As for T_2, all the users but User u_1 consent to a more privacy-invasive task (e.g., displaying the noise measurements across the users' journeys). Assigning tasks to users based only on their respective task consents (i.e., the system assigns T_1, resp. T_2, to all 10, resp. 9, users) results in a defect in consent. Indeed, although User u_1 does not consent to reveal the detailed noise observations they contribute to, the specific observations may be inferred from composing O_{T_1} and O_{T_2}.

User	b_1	b_2	b_3	Hilbert
u_1	1	0	0	1
u_3	1	1	1	5
u_4	1	1	1	5
u_6	1	1	1	5
u_7	1	1	1	5
u_9	1	1	1	5
u_{10}	1	1	1	5
u_2	1	1	0	6
u_5	1	1	0	6
u_8	1	1	0	6

TABLE I: Consent table $\mathcal{C}$ (see § IV for Hilbert column).

The consent-driven task assignment must account for the disclosure policies of the users across all the tasks they each contribute to, to avoid a defect in consent. In other words, the multi-task assignment must be achieved in such a way that the data gathered by a task T_1 cannot be analyzed together with data also gathered by a task T_2 with a weaker disclosure policy. A simple solution to the above issue would consist in assigning tasks according to either the least or the greatest, common disclosure policy of the eligible users. Going back to our example, this means assigning task T_1 to the participants consenting to either T_1 only, or both T_1 and T_2. However, this would lead to sub-optimal data reuse, with a contribution loss for certain tasks and a resulting reduced utility. The research question that the paper addresses is then: *How to avoid by design the defects in consent in a multi-tasking participatory sensing system, while ensuring efficient data reuse in as many eligible tasks as possible?*

III. Consent-driven ℓ-Completeness for Privacy-aware Data Reuse

Consider the set $\mathcal{T} = \{T_i\}_{0<i\leq m}$ with any $T_i =< f_i, \mathcal{S}_i, \Delta, \ell >$ (values of ℓ and Δ are identical in all tasks for simplicity, but without loss of generality –see § II). It is direct to infer that a multi-task assignment does not create any defect in consent if for any pair of distinct tasks T_i and T_j, their respective outputs $O_{T_i} = f_i(s_{U_i})$ and $O_{T_j} = f_j(s_{U_j})$ are processed over the respective input sets S_{U_i} ($|s_{U_i}| \geq \ell$) and S_{U_j} ($|S_{U_j}| \geq \ell$) such that either $S_{U_i} = S_{U_j}$ (i.e., they analyze the very same set of contributions from the same **consenting** users) or $S_{U_i} \cap S_{U_j} = \emptyset$ (i.e., they analyze contributions from distinct sets of users). This leads us to introduce the following strict definition of the $\ell-$*Completeness* property:

Definition 2 (**Strict $\ell-$Completeness over Δ**). *Given a set of m tasks, $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$, and a set of registered users, $\mathcal{U}$, who consent to (contribute to specified subsets of) these tasks and thereby produce input data for these tasks over a time period Δ, a family of computation input sets $(S_{U_i})_{i\leq m}$, where U_i is the set of users contributing (and obviously consenting) to $T_i \in \mathcal{T}$, is $\ell-$complete over Δ iff:*

- *$\forall i \leq m$: $|U_i| \geq \ell$*
- *$\forall i, j \leq m$: either $U_i \cap U_j = \emptyset$ or $U_i = U_j$*

As shown in Fig. 1, the above definition suits well tasks that focus on mutually disjoint (see *no reuse* scheme in Fig. 1) or fully identical (see *strict reuse*) sets of potentially usable contributions. However, given a set of users U_1 consenting (only) to a task T_1 and a (disjoint) set of users U_{12} consenting to tasks T_1 and T_2, if T_1 uses all usable contributions (i.e., the result $O_{T_1} = f_1(S_{U_1} \cup S_{U_{12}})$ is produced) then Definition 2 precludes T_2 from using any contribution in $S_{U_{12}}$ produced by any user in U_{12} (i.e., only result $O_{T_2} = f_2(S_{U_2})$ can be produced, but not $O_{T_2} = f_2(S_{U_{12}})$ neither $O_{T_2} = f_2(S_{U_2} \cup S_{U_{12}})$).

To enable a better data reuse in practice, we propose a relaxation of Definition 2, while still prohibiting any defect in consent. This leads to Definition 3:

[3] Solutions under lower security assumptions, e.g., semi-honest system or covert adversary, are considered future work.

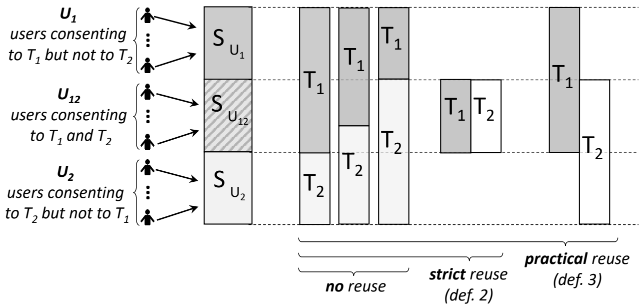

U_1
users consenting to T_1 but not to T_2
U_{12}
users consenting to T_1 and T_2
U_2
users consenting to T_2 but not to T_1
S_{U_1}
$S_{U_{12}}$
S_{U_2}
T_1
T_2
no reuse
strict reuse (def. 2)
practical reuse (def. 3)

Fig. 1: Strict vs. practical $\ell-$complete data reuse schemes.

Definition 3 (**Practical $\ell-$Completeness over Δ**). *Let a set of m tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ and a set of users $\mathcal{U}$ who consent to (a desired subset of) these tasks and contribute input data for these tasks over a period of time Δ. Let a family of computation input sets $(S_{U_i})_{i \leq m}$, where U_i is the set of users consenting to $T_i \in \mathcal{T}$ whose contribution is used in $(S_{U_i})_{i \leq m}$ to evaluate tasks $T_i \in \mathcal{T}$. Let $F(\mathcal{X}) = \{X \cap X' | X, X' \in \mathcal{X}\} \bigcup \{X \setminus X' | X, X' \in \mathcal{X}\} \bigcup \mathcal{X}$. Let $\mathcal{F} = \bigcup_{k \in \mathbb{N}} F^k(\{U_i\}_{i \leq m})$ be the least fixed point of F containing $\{U_i\}_{i \leq m}$.*

The family of computation input sets $(S_{U_i})_{i \leq m}$ is $\ell-$complete over Δ iff for all $X \in \mathcal{F}$ either $X = \emptyset$ or $|X| \geq \ell$.

In addition to the *no reuse* and *strict reuse* schemes, Definition 3 allows for further data reuse (see *practical reuse* in Fig. 1). Reuse is enabled as long as applying any combination of the intersection and difference over the respective users' contributed data consumed by different tasks does not produce information on a set of less than ℓ users' contributions. In particular, Definition 3 allows evaluating both $O_{T1} = f_1(S_{U_1} \cup S_{U_{12}})$ and $O_{T_2} = f_2(S_{U_2} \cup S_{U_{12}})$ when the conditions of practical $\ell-$completeness are met. The following sections introduce concrete techniques for data reuse, which conform by design to Definition 3.

IV. Task Assignment Design and Evaluation

We aim at eliciting groups of ℓ users that foster consent-driven data reuse across tasks according to Definition 3. To achieve so, we leverage the consent table $\mathcal{C}$ (e.g., see Table I).

A. ℓ-completeness as a partitioning problem

The simplest approach to group users using the consent table is to create partitions that group together users specifying identical policies. Then, if the size of the partition is more than ℓ, the users of the partition are assigned all the tasks associated with the corresponding consent policy. Otherwise, the partition is ignored. This approach is well suited to involve users sharing a consent policy with a large number of peers. However, it tends to set aside users consenting to a policy that differs (whether in a weaker or stronger form) from that of most users, which may result in a high contribution loss. Specifically, given m tasks and n users, if every task requires ℓ users, then there are $(C_n^\ell)^m$ different combinations, each defining a potential distinct consent policy. A possible countermeasure would be to incentivize users to change their consent (e.g., according to the "privacy for reward" of auction-based approaches [19], [20]). However, we consider that this significantly alters privacy and that the users' specified consents are an immutable parameter.

The $\ell-$completeness problem can then be efficiently solved by sorting and partitioning the consent table such that each partition contains exactly ℓ users. An approach is to transform the multidimensional space into a one-dimension space, while preserving data locality, over which a partitioning algorithm allows creating groups of ℓ participants. The *Space filling curve* techniques are good candidates for such transformation, among which the *Hilbert space filling-curve* offers the best results [38]. That is, we use the Hilbert curve to map the consents to multiple tasks of each user to a single index characterized by its Hilbert value. We may then sort the users according to their associated consent-related Hilbert values. Following, we assign the first ℓ users (as ordered according to the Hilbert value) to the first partition, the next ℓ users to the second partition, and so on. Consequently, each partition contains exactly ℓ users (except for the last one that may contain up to $2\ell - 1$ users). As a result, each partition groups users who are close with respect to their consent policies, as abstracted using the Hilbert curve. Finally, an adjustment phase allows discarding the tasks associated with the partition (as defined by the consents of the embedded users), to which at least one user does not consent so that the users of a given partition are all assigned the same set of tasks to which they all consented.

However, the above approach does not systematically lead to group the users that are the closest with respect to their respective consent policies. Consider the consent Table I and $\ell = 5$ with the rightmost column providing the associated Hilbert values (precision=3). Applying the proposed algorithm results in two partitions: $P_1 = \{u_1, u_3, u_4, u_6, u_7\}$ and $P_2 = \{u_9, u_{10}, u_2, u_5, u_8\}$. The corresponding greatest common consent policy is $< 1, 0, 0 >$ in P_1 and $< 1, 1, 0 >$ in P_2, leading to a loss of 10 ($2{*}\ell$) contributions. A more optimal partitioning would be: $P_1 = \{u_3, u_4, u_6, u_7, u_9\}$ with a major consent policy $< 1, 1, 1 >$ and $P_2 = \{u_1, u_2, u_5, u_8, u_{10}\}$ with a major consent policy $< 1, 0, 0 >$, resulting in a loss of ℓ contributions. That is, the partitioning algorithm needs to prioritize the grouping of users who are similar consent-wise, so that the greatest common consent policy within a group maximizes the number of eligible users.

B. ℓ-completeness as a clustering problem

The key idea underlying our approach is to consider the $\ell-$completeness problem as a clustering problem. Clustering is defined as partitioning a set of objects into groups such that objects in the same group are more similar to each other than objects in other groups with respect to predefined distance criteria. The $\ell-$completeness problem can then be formulated as identifying a set of clusters such that each cluster contains at least ℓ users while minimizing the overall contribution loss.

We thus propose to group users based on the similarity of their consent policies, and assigning them the tasks that match the consent policies of all, while discarding the others. Hence, this allows users with singular policies to contribute.

The objective is to create a set of k clusters $\mathcal{P} = \{P_1, P_2, ...P_k\}$ of the consent table $\mathcal{C}$ in such a way that the tuples (i.e., the consent policies of the embedded users) in the same cluster are as similar to each other as possible, thus resulting in a minimum contribution loss. That is, we want to minimize the sum of all the intra-cluster distances (maximum distance between any two data points), which is defined as:

$$d = \sum_{h=1...k} \max_{i,j=1,...,|P_h|} distance(x_{h,i}, x_{h,j}) \qquad (1)$$

where: $x_{h,i}$ denotes the i^{th} data point (in our case, a consent policy of the form $< b_1, \ldots, b_m >$) in the cluster P_h of the consent table $\mathcal{C}$, and *distance(x,y)* is the Euclidean distance between two data points x and y and serves characterizing the dissimilarity (in terms of common consents) between users. It follows that we aim at eliciting k subsets of users, P_i $(1 \leq i \leq k)$, such that:

- $\forall i \neq j \in \{1...k\} : P_i \cap P_j = \emptyset$
- $\forall i \in \{1...k\} : |P_i| \geq \ell$
- The distance d (defined in Equation 1) is minimized.

The supporting algorithm subdivides into 3 stages: (1) *Clustering* that assigns users to their respective clusters while maximizing the number of tasks to be executed by each user; (2) *Adjustment* to deal with clusters that contain less than ℓ users and further enforce the single greatest common consent policy, and (3) *Selection* so that the task assignment to clusters maximizes data reuse.

1) Clustering stage: The $\ell-$completeness problem does not have a constraint on the number of clusters; however, it requires that each cluster contains at least ℓ users. Thus, we pose the $\ell-$completeness problem as a clustering problem that derives from the traditional $k-$means algorithm. We specifically choose $k-$means because it is one of the most widely used algorithms for clustering due to its simplicity and efficiency with a low computational overhead [39]. Still, to overcome the limits of the standard $k-$means algorithm for choosing the k initial random centers, we use the combined $k-$means^{++} that allows achieving better accuracy by choosing the starting centers based on the weights of the data points according to their squared distance from the closest center already chosen [40]. That is, we use $k-$means^{++} to seed the initial centers for $k-$means in such a way that they are as far apart from each other as possible.

Algorithm 1 Similarity clustering

```
Input: A consent table $\mathcal{C}$, the value $\ell$ for $\ell-$completeness
Output: A set of clusters $\{P_1, ...P_k\}$
  Let $k \leftarrow \left\lfloor \frac{n}{\ell} \right\rfloor$
  Let $\{P_1, ..., P_k\}$ a set of $k$ empty clusters
  Select $k$ distinct tuples $c_1...c_k \in \mathcal{C}$ with $k-$means$^{++}$
  $\mathcal{X} \leftarrow \{c_1, ..., c_k\}$        $\triangleright$ $\mathcal{X}$ contains the initial $k$ centroids
  repeat
      for each tuple c $\in \mathcal{C}$ do
          Find $c_i \in \mathcal{X}$ closest to $c$ using Equation 1
          $P_i \leftarrow P_i \cup \{c\}$
      end for
      $\mathcal{X} \leftarrow \{centroid(P_i)\}_{i \leq k}$
  until convergence
  return $\{P_1, ..., P_k\}$
```

This results in Algorithm 1 that proceeds as follows: Let $\mathcal{C}$ be the consent table and $k = \left\lfloor \frac{n}{\ell} \right\rfloor$ such that n is the number of users in table $\mathcal{C}$ and ℓ is the value set for the $\ell-$completeness. The clustering stage starts by selecting k tuples to build k clusters using $k-$means^{++}. The idea of $k-$means^{++} is to choose a random consent c_1 from $\mathcal{C}$, calculate the distance from each data point to the closest center we have already chosen, sample a point with a probability proportional to the square of the distance already calculated and repeat the previous two steps until k centroids are selected. Once the initial k centers are chosen, the $k-$means algorithm is applied: for each tuple c in the consent table $\mathcal{C}$, the algorithm finds the cluster P_i with the closest centroid to c. Then, we add c to its closest cluster and subsequently update the centroid of the clusters.

2) Adjustment stage: The above clustering algorithm returns a set of clusters out of which some may contain less than ℓ users. Furthermore, it is unlikely that the consent policies of all the users in a given cluster are identical, which requires defining the *Greatest Common Policy* (*GCP*) for the cluster:

Definition 4. ***(Greatest Common Policy – GCP)*** *Given a cluster $P_i \subset \mathcal{C}$, we define the Greatest Common Policy within P_i as the tuple $GCP(P_i) =< b_1, b_2, .., b_m >$ such that $b_i = 1$ if for any tuple p in P_i we have $p.b_i = 1$, and $b_i = 0$ otherwise.*

That is the GCP of the cluster characterizes the allowable set of tasks so as to avoid any defect in consent (see § II). Algorithm 2 introduces the computation of the clustering adjustment. This includes checking the number of users in each cluster. Multiple approaches can be used in case $|P_i| < \ell$: (1) Distribute all the tuples of the cluster to the closest clusters; (2) Merge closest small clusters with P_i until $|P_i| \geq \ell$; (3) Discard Cluster P_i. For efficiency and simplicity, we implement Option 1.

3) Selection stage: The objective of this third stage is to assign the tasks to clusters, while optimizing data reuse. The clustering stage defines the consent policy for the cluster of users in terms of the set of allowable tasks. For each task T_i, we merge all clusters P_j such that $b_i = 1$ in $GCP(P_j)$ in order to maximize data reuse. Note that some of the clusters may be reused in many tasks.

Compliance with practical $\ell-$completeness. The Clustering Algorithm 1 and associated Adjustment Algorithm 2 build a partition of the users' contributions such that each partition is a cluster of size $\geq \ell$ and the intersection of any two partitions is the empty set. Since each task (re-)uses the union of a given set of clusters, any composition of union, intersection and dif-

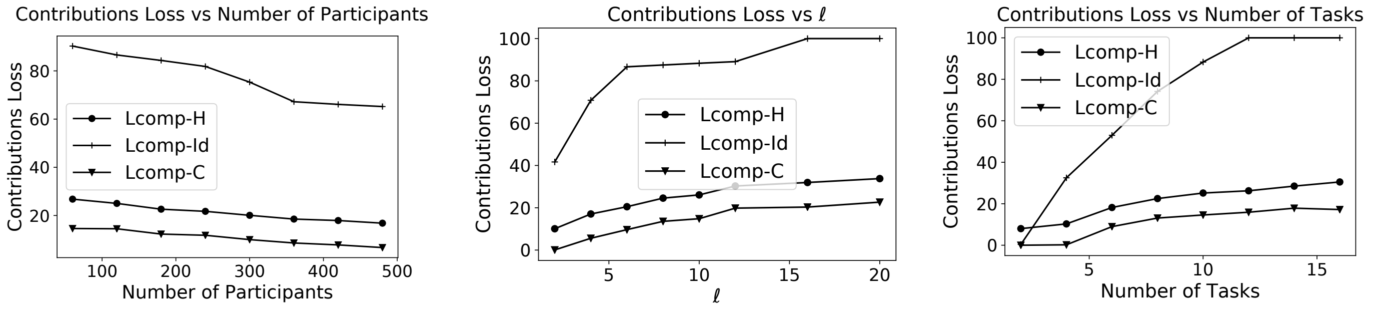


Fig. 2: Partitioning vs clustering for task assignment: *Lcomp-C* refers to our clustering algorithm based on Algorithms 1 and 2, *Lcomp-H* refers to partitioning based on Hilbert values, and *Lcomp-Id* refers to partitioning based on identical consents.

Algorithm 2 Clusters adjustment

```
Input: $\mathcal{P} = \{P_i\}_{i \leq k}$ a set of $k$ clusters
Output: $\{P_i{}'\}_{i \leq k'}$ a set of $k' \leq k$ clusters with $|P_i{}'| \geq \ell$
  for each cluster $P_i \in \mathcal{P}$ with $|P_i| < \ell$ do
    for each  tuple $x \in P_i$  do
      Find $P_j$ with $centroid(P_j)$ closest to $x$ and $|P_j| \geq \ell$
      $P_j \leftarrow P_j \cup \{x\}$
      $P_i \leftarrow P_i \setminus \{x\}$
    end for
    $\mathcal{P} \leftarrow \mathcal{P} \setminus \{P_i\}$
  end for
  return $\mathcal{P}$
```

ference over the sets of users whose contribution is effectively used in the different tasks is obviously either the empty set or a set containing one or more partitions (corresponding to more than ℓ users by construction). Hence, the result complies with practical $\ell-$completeness (see Definition 3) by design.

C. Contribution loss

The effectiveness of the proposed clustering algorithm relates to the extent to which it fosters the reuse of data across tasks, as allowed by the consents of the contributing users. Using a consent table filled according to a Bernoulli distribution (with a probability $\alpha = 0.6$), we show below that the clustering approach (§ IV-B) outperforms base partitioning (§ IV-A). Others distributions types could be used, for example a Zipf distribution.

$$C_{loss}(\{T_i\}_{i \leq m}) = \frac{|\{c.b_i \in \mathcal{C}, i \leq m, c.b_i = 1, GCP(c).b_i = 0\}|}{|\{c.b_i \in \mathcal{C}, c.b_i = 1\}|} \tag{2}$$

Equation 2 defines the ratio of contributions loss within a purely consent- driven system.[4] Fig. 2 reports the total contributions loss that represents the number of contributions that are discarded by the adjustment stage divided by the total number of contributions (see Equation 2) for the three algorithms according to: (*Left*) Number of participants with 12 tasks and $\ell = 10$; (*Center*) Value of ℓ with 12 tasks and 120 participants; and (*Right*) Number of tasks with 120 participants and $\ell = 10$. As depicted, our clustering algorithm results in the least loss in terms of contributions for all the evaluations. It also shows a significant loss of contributions for the first partitioning algorithm that is based on identical consents, which is due to the fact that the majority of the groups are smaller than ℓ. We observe (Left) that the information loss diminishes with the increase in the number of participants using our clustering algorithm; the reason is that increasing the number of participants creates more users with similar consents. Thus, the clustering algorithm becomes more efficient as more users have similar consents. This property also holds for the partitioning approach based on Hilbert values. Still, the clustering algorithm offers the best performance. We notice that for the partitioning based on identical consents, the loss is total by increasing the value of ℓ or the number of tasks (Center). This means that the algorithm fails to identify groups of users with more than ℓ identical consents. The superiority of our clustering approach results from the fact that it creates groups of at least ℓ while using a distance function that measures the similarities between consents. We also leverage an optimized algorithm for the selection of starting centroids, leading to reduce the information loss in the adjustment phase. We performed additional experiments to investigate the effect of varying α on our results and we found that the curves maintain a similar shape with a variation in the slope. The loss of contributions is higher when decreasing α and it is lower when increasing α. The reason is that decreasing α generates more users who have not consented to tasks, thus choosing a GCP for each cluster discards more consents.

D. Utility loss

In addition to minimizing the contributions loss, the clustering strategy must also optimize the data utility. That is, we need to assess the effectiveness of our clustering algorithm in leveraging the utility of the contributed data to the assigned tasks. We analyze the effectiveness of our algorithm against data utility using a dataset made available to us by the authors of [41], which provides the users' activities, positions and environmental noise measurements collected by their crowdsensing app. We specifically use a dataset containing $47,954$

[4] Non consent-driven approaches are out of the scope of this paper.

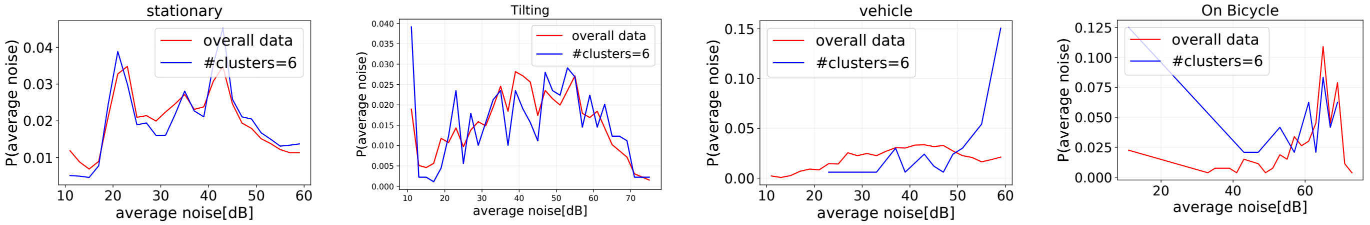


Fig. 3: Distribution of the collected values per activity after clustering

data points, collected by 120 participants over a period of one week. We further carry out the analysis considering an illustrative task that computes the average daily exposure to noise according to the user's activity (i.e., Stationary, Walking, In-vehicle, Tilting, On-bicycle).

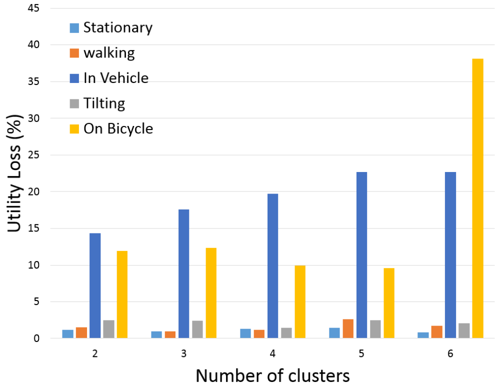


Fig. 4: Utility loss per activity *wrt* number of clusters.

For any result O_{T_i} obtained for a task T_i, we evaluate the utility loss as follows:

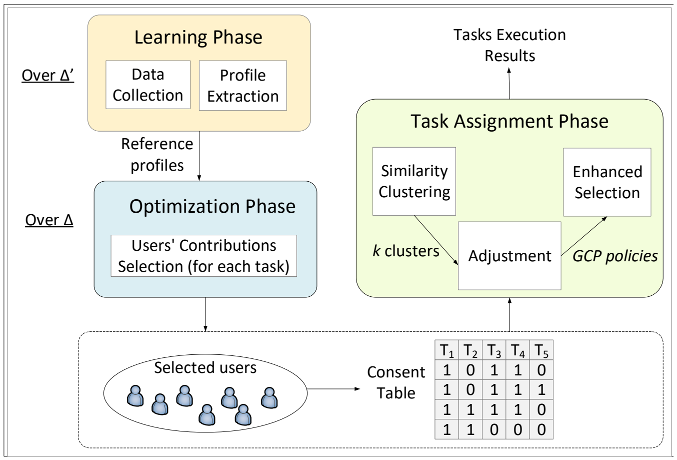
$$U_{loss}(T_i) = \frac{f_i(\{S_U | C_U \in P_i, GCP(P_i).b_i = 1\})}{f_i(\{S_U | C_U.b_i = 1\})} \quad (3)$$

with S_U data contributed by a set of users U. In the case of a multi-dimensional result (e.g., noise values aggregated by activity), we evaluate the value of U_{Loss} for each dimension.

Fig. 4 analyzes the impact of our algorithm on the utility loss (see Equation 3) by comparing the execution of the task on the clusters (with clustering) with the execution on the original data (without clustering), according to the number of clusters $|\mathcal{P}| \leq k$ with $k = \left\lfloor \frac{n}{\ell} \right\rfloor$ (see Algorithm 1). We varied the ℓ value for this evaluation. However, investigating the optimal ℓ value is area for future work, and depends on the set of tasks and the related consents. As depicted, the information loss tends to be larger when we increase the number of clusters for certain types of activities (On-bicycle, In-vehicle), while clustering does not have any significant impact on the data utility for others (Stationary, Tilting). In general, if the cluster is a representative sample of the target observations, the information loss is limited. That is, the utility loss is directly proportional to the difference between the data distribution within the clusters and the overall set of contributory data. Then, if the distance between the distribution of the target attribute (e.g., Noise exposure) in the clusters and the distribution in the whole table is limited, the information loss is limited. To motivate this intuition, Fig. 3 compares the distribution of data in the clusters with the one of the overall dataset. Thus, the difference in the probability distributions assesses the utility loss: The more the distributions of data are different, the more the information loss is important. Our clustering algorithm does not account for such a criterion and thus impacts the data utility resulting from the proposed consent-driven multi-task assignment, which we overcome next.

V. Enhancing the ℓ-Completeness Algorithm for Higher Accuracy

We enhance the utility of our consent-driven clustering process by evaluating a data distribution profile for each task beforehand, which allows us to optimize the assignment of data to clusters. This results in the following three phases for the multi-task assignment (see Fig. 5).



T_1	T_2	T_3	T_4	T_5
1	0	1	1	0
1	0	1	1	1
1	1	1	1	0
1	1	0	0	0

Fig. 5: Three-phase consent-driven multi-task assignment.

A. Learning phase

The learning phase runs over a time period $\Delta' < \Delta$ –which is considered representative– to generate, for each task T, the *reference profile* φ_0^T of the data distribution associated with the task. For example, for tasks related to urban monitoring, the typical value for Δ' would be a week as we observe weekly repetitive patterns. We do not detail the computation of the reference profiles, which are task dependent. We simply highlight that the learning phase does comply with the users' consent policies because: (i) The system computes every profile φ_0^T using a set of clusters $\mathcal{P}$ such that $\forall P \in \mathcal{P} : GCP(P).b_i = 1$ (which guarantees $\ell-$completeness over Δ'); and (ii) The system does not (directly) expose the profiles, which are only used to construct $\ell-$complete input sets for subsequent periods.

B. Optimization phase

The second phase leverages the set of reference profiles $\{\varphi_0^T\}_{T\in\mathcal{T}}$ to optimize the users contributions to be considered for each task for time period Δ. For each task T, we compute the profile φ_Δ^T as the union of the contributions of all the users consenting to T over Δ. We then compare φ_Δ^T with T's reference profile φ_0^T; if the two profiles diverge, we mark the contributions of some users as *'ignored'* for this task (i.e., the corresponding value b_i in the consent table is updated from 1 to 0) until φ_Δ^T gets close enough to φ_0^T. The way users' contributions are selected is task/profile dependent (e.g., if the profile is a distribution or an histogram, the contributions of users with over represented values may be 'ignored') and is not further discussed. Note that the contributions of users that get 'ignored' for one task can still be considered for any other –consented– task. The optimization phase allows "ignoring" subsets of users contributions that would otherwise negatively impact the data utility, while confirming to the users' privacy and consent requirements through the computation of $\ell-$complete clusters.

C. Enhanced selection phase

While the optimization phase filters out users contributions so as to respect the task profile φ_Δ^T of any task T, the clustering algorithm (Algorithms 1 & 2, § IV) may still introduce some bias since it leads to use only a subset of the eligible data for the task. Algorithm 3 overcomes such an impact by enhancing the selection defined in § IV-B3. Given a set of clusters $\mathcal{P}$ created after proportional clustering/adjustment and a set of task profiles $\{\varphi_\Delta^T\}_{T\in\mathcal{T}}$, for each task T, Algorithm 3 starts by merging the set of clusters $P \in \mathcal{P}$ with $T \in GCP(P)$ in a cluster E_T. Then, it computes the profile $\varphi_\Delta^T(E_T)$ on E_T for that task. If the resulting profile is close enough to the profile φ_Δ^T, E_T is considered as the input set for T that may be used to compute task result O_T. Otherwise, we compute the profile of each cluster in E_T, sort them according their distance to φ_Δ^T and we remove those with the highest EMD distance until finding a qualified cluster.

More precisely, to quantify the difference between data distributions profile in the clusters and data distribution profile φ_Δ^T, we use the Earth Mover's Distance (EMD) [42]. The EMD between two distributions R and S is defined as:

$$distance(R,S) = \frac{1}{m-1}\sum_{i=1}^{m-1}|\sum_{i=j}^{i}(s_j - r_i)|$$

As a result, Algorithm 3 selects, for each task, an input set that maximizes data reuse and respects the task profile.

Algorithm 3 Enhanced selection

```
Input: $\mathcal{P}$: the set of clusters created after clustering/adjustment, $\{\varphi_\Delta^T\}$: the set of task profiles
Output: The set of input sets $\{E_T\}$ qualified for each task
  for each $T \in \mathcal{T}$ do
      $E_T \leftarrow \cup\{P \in \mathcal{P}, T \in GCP(P)\}$
      qualified ← false
      while ! qualified do
          $\varphi_{E_T} \leftarrow compute\_profile_T(E_T)$
          if $\varphi_{E_T} \simeq \varphi_\Delta^T$ then
              qualified ← true
          else          ▷ else remove the worst cluster from $E_T$
              $Worst \leftarrow P \in E_T$ s.t. $P$ with the highest
                  $EMD(compute\_profile_T(P), \varphi_\Delta^T)$
              $E_T = E_T \setminus Worst$
          end if
      end while
  end for
  return $\{E_T\}$
```

D. Utility Loss

Fig. 6 illustrates the effectiveness of our enhanced selection stage to achieve higher utility and less information loss compared to our original clustering algorithm (see Fig. 4). Specifically, it shows the information loss when measuring the noise pollution level. We compare the execution on the overall data *vs* a qualified cluster (i.e., the input set produced with Algorithm 3 for that task). We see that the information loss remains negligible with the increase in the number of clusters. The reason is that the cluster follows the global distribution of data; therefore, the difference between the distribution of data in the cluster and the original data is low, which results in a low information loss.

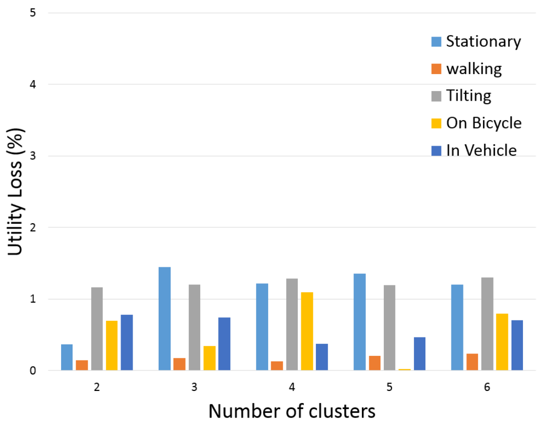


Fig. 6: Utility Loss per activity *wrt* # of clusters.

VI. Related Work

Multi-task allocation is a key research issue in participatory sensing systems and has attracted much attention from researchers in recent years.

A. Multi-task allocation in participatory sensing systems

The goal of the multi-task allocation is to minimize a total cost function while guaranteeing data quality for multiple tasks. In [43], the objective is to select the minimum subset of participants that satisfies quality-of-information metrics (i.e., granularity and quantity) under a total budget constraint. Zhang *et al.* [44] introduce a strategy to predict the mobility of participants so as to select the minimum number of participants while ensuring the best spatio-temporal coverage. Li *et al.* [45] also aim at minimizing the number of participants while meeting a predefined level of coverage. *TaskMe* [46], a framework for multi-task allocation, deals with (1) maximizing the number of accomplished tasks when few participants are available; and (2) the opposite problem with many participants and few tasks. *PSAllocator* [24] addresses the multi-task allocation problem from another perspective, that is, maximizing the overall system utility under sensing capability constraints. This includes the work in [25] that aims at maximizing data utility according to participant-side factors (e.g., participant bandwidth, participant availability) when assigning tasks. Zhu *et al.* [47] introduce a greedy-based approach to maximize the number of accomplished tasks under the sensing capacity and time constraints.

In general, existing solutions to the multi-task assignment mainly focus on optimizing the allocation of tasks without taking into account the consents of users. Our work thus distinguishes itself by aiming at maximizing data reuse while taking into account users' consents, which is essential when dealing with the collection of users' data as with crowdsensing systems. In addition, by considering the distribution of data, the overall system utility is optimized.

B. Consent in computing platforms

To the best of our knowledge and despite the growing awareness about privacy and consent in multiple fields, the concept of consent has never been considered in the existing literature about mobile crowdsensing systems. Yet, Luger *et al.* [48] alert multidisciplinary experts about *"a crisis of consent for ubiquitous computing"*. They call the designers to balance their design objectives against a series of consent considerations. Jones [26] discusses the importance of consent in computing, explaining the moral magic of consent that renders permissible an otherwise impermissible action. In [49], [50], the authors study the role of consent in privacy policies for social media users. In the case of Facebook, they consider that consent is flawed and claim the need for improvement to create more transparency about users' personal data. Recently, Okoyomon *et al.* [51] compare the privacy policies of Google Play Store apps with their behaviors and highlight the level of the defect in consent and lack of prioritizing user privacy. In light of this, our work introduces a consent-driven approach to multi-task allocation in participatory sensing systems. Our solution specifically maximizes the number of tasks assigned to participants according to their consents while minimizing the utility loss.

VII. Conclusion

We have introduced the $\ell-$completeness consent-driven property for multi-task allocation in mobile crowdsensing systems so as to foster the reuse of crowdsensed contributions across eligible tasks while strictly adhering to the consents of users. Indeed, respecting *consent* is an essential property of crowdsensing systems to guarantee privacy to their users. This leads us to claim that consent-based properties, such as $\ell-$completeness, should be an essential part of standards oriented towards privacy in crowdsensing systems.

We have also presented a solution to multi-task allocation that conforms to $\ell-$completeness, while optimizing data reuse and utility. Initial evaluation using a dataset from a deployed crowdsensing app shows the relevance of the supporting algorithms.

Our current and future work concentrate on addressing a more complex attacker model by assuming a malicious adversary that may gain root access to the crowd-sensing system, modify data or eavesdrop communications. Our future objective is to ensure the integrity and confidentiality of the proposed solution through the use of trusted execution environments (e.g., Intel SGX enclaves). That is, this will guarantee that no user's contribution can ever be exposed in clear text outside secure enclaves and apart from the tasks results produced, which we will implement within a system for validation.

VIII. Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work is partially funded by the GDP-ERE project of the Convergence Institute DATAIA of the University of Paris Saclay.

References

[1] J. Liu, H. Shen, H. S. Narman, W. Chung, and Z. Lin, "A survey of mobile crowdsensing techniques: A critical component for the internet of things," vol. 2, no. 3, 2018.

[2] A. Longo, M. A. Bochicchio, and M. Zappatore, "Apollon project: A massive online open lab for citizen science driven environmental monitoring," in *2020 IEEE Global Engineering Education Conference (EDUCON)*, 2020, pp. 1703–1712.

[3] L. A. Kalogiros, K. Lagouvardos, S. Nikoletseas, N. Papadopoulos, and P. Tzamalis, "Allergymap: A hybrid mHealth mobile crowdsensing system for allergic diseases epidemiology : a multidisciplinary case study," in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2018, pp. 597–602.

[4] J. Wan, J. Liu, Z. Shao, A. Vasilakos, M. Imran, and K. Zhou, "Mobile crowd sensing for traffic prediction in internet of vehicles," *Sensors*, vol. 16, no. 1, Jan 2016.

[5] X. Zhang, Z. Yang, W. Sun, Y. Liu, S. Tang, K. Xing, and X. Mao, "Incentives for mobile crowd sensing: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 54–67, 2016.

[6] D. Wu, S. Si, S. Wu, and R. Wang, "Dynamic trust relationships aware data privacy protection in mobile crowd-sensing," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2958–2970, 2018.

[7] X. Hu, X. Li, E. C. . Ngai, V. C. M. Leung, and P. Kruchten, "Multidimensional context-aware social network architecture for mobile crowdsensing," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 78–87, 2014.

[8] Y. Du, F. Sailhan, and V. Issarny, "Let opportunistic crowdsensors work together for resource-efficient, quality-aware observations," in *PerCom 2020: IEEE International Conference on Pervasive Computing and Communications*, 2020.

[9] D. Christin, "Privacy in mobile participatory sensing: Current trends and future challenges," *Journal of Systems and Software*, vol. 116, pp. 57 – 68, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0164121215000692

[10] I. Krontiris, M. Langheinrich, and K. Shilton, "Trust and privacy in mobile experience sharing: future challenges and avenues for research," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 50–55, 2014.

[11] L. Pournajaf, D. A. Garcia-Ulloa, L. Xiong, and V. Sunderam, "Participant privacy in mobile crowd sensing task management: A survey of methods and challenges," *SIGMOD Rec.*, vol. 44, no. 4, p. 23–34, May 2016.

[12] M. A. Alsheikh, Y. Jiao, D. Niyato, P. Wang, D. Leong, and Z. Han, "The accuracy-privacy trade-off of mobile crowdsensing," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 132–139, 2017.

[13] L. Sweeney, "K-anonymity: A model for protecting privacy," vol. 10, no. 5, 2002.

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 24–24.

[15] L. Kazemi and C. Shahabi, "A privacy-aware framework for participatory sensing," *SIGKDD Explor. Newsl.*, vol. 13, no. 1, p. 43–51, Aug. 2011.

[16] L. Wang, D. Zhang, D. Yang, B. Y. Lim, X. Han, and X. Ma, "Sparse mobile crowdsensing with differential and distortion location privacy," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2735–2749, 2020.

[17] C. Luo, X. Liu, W. Xue, Y. Shen, J. Li, W. Hu, and A. X. Liu, "Predictable privacy-preserving mobile crowd sensing: A tale of two roles," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 361–374, 2019.

[18] J. Lin, D. Yang, M. Li, J. Xu, and G. Xue, "Frameworks for privacy-preserving mobile crowdsensing incentive mechanisms," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1851–1864, 2018.

[19] H. Jin, L. Su, H. Xiao, and K. Nahrstedt, "Incentive mechanism for privacy-aware data aggregation in mobile crowd sensing systems," *IEEE/ACM Transactions on Networking*, vol. 26, no. 5, pp. 2019–2032, 2018.

[20] Z. Wang, J. Hu, R. Lv, J. Wei, Q. Wang, D. Yang, and H. Qi, "Personalized privacy-preserving task allocation for mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1330–1341, 2019.

[21] J. Ni, K. Zhang, Q. Xia, X. Lin, and X. S. Shen, "Enabling strong privacy preservation and accurate task allocation for mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1317–1331, 2020.

[22] C. Zhao, S. Yang, and J. A. McCann, "On the data quality in privacy-preserving mobile crowdsensing systems with untruthful reporting," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2019.

[23] K. Mišura and M. Žagar, "Data marketplace for internet of things," in *2016 International Conference on Smart Systems and Technologies (SST)*, 2016, pp. 255–260.

[24] J. Wang, Y. Wang, D. Zhang, F. Wang, Y. He, and L. Ma, "PSAllocator: Multi-task allocation for participatory sensing with sensing capability constraints," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 1139–1151.

[25] J. Wang, F. Wang, Y. Wang, D. Zhang, B. Lim, and L. Wang, "Allocating heterogeneous tasks in participatory sensing with diverse participant-side factors," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 1979–1991, 2018.

[26] M. L. Jones, "The development of consent to computing," *IEEE Annals of the History of Computing*, vol. 41, no. 4, pp. 34–47, 2019.

[27] J. E. Bartlett, J. W. Kotrlik, and C. C. Higgins, "Organizational research: Determining appropriate sample size in survey research," *Information Technology, Learning, and Performance Journal*, vol. 19, 2001.

[28] C. M. O'Keefe, "Privacy and confidentiality in service science and big data analytics," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2014, pp. 54–70.

[29] European Council, "Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46," *Official Journal of the European Union (OJ)*, vol. 59, no. 1-88, p. 294, 2016.

[30] S. L. Pardau, "The California Consumer Privacy Act: Towards a European-Style Privacy Regime in the United States," *J. Tech. L. & Pol'y*, vol. 23, p. 68, 2018.

[31] "European Commission explanations," https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/purpose-data-processing/can-we-use-data-another-purpose_en.

[32] M. Fabre-Magnan, *Defects on Consent in Contract Law*. European Contract Code, 2. Ed., The Hague 1998, 2004.

[33] R. Ventura, V. Mallet, and V. Issarny, "Assimilation of mobile phone measurements for noise mapping of a neighborhood," *Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1279 – 1292, Sep. 2018.

[34] French National Institute of Statistics and Economic Studies (INSEE), "Guide to statistical confidentiality," Tech. Rep., 2020. [Online]. Available: https://www.insee.fr/en/statistiques/fichier/2388575/guide-secret-en.pdf

[35] Secure Access Data Center (CASD), "Confidentiality rules," Tech. Rep., 2020. [Online]. Available: https://www.casd.eu/wp/wp-content/uploads/Output_Confidentiality_Rules.pdf

[36] "anonytl specification."

[37] M. Shin, C. Cornelius, D. Peebles, A. Kapadia, D. Kotz, and N. Triandopoulos, "AnonySense: A system for anonymous opportunistic sensing," *Pervasive and Mobile Computing*, vol. 7, no. 1, pp. 16–30, 2011.

[38] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz, "Analysis of the clustering properties of the Hilbert space-filling curve," *IEEE Transactions on knowledge and data engineering*, vol. 13, no. 1, pp. 124–141, 2001.

[39] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[40] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.

[41] V. Issarny, V. Mallet, K. Nguyen, P.-G. Raverdy, F. Rebhi, and R. Ventura, "Dos and Don'ts in Mobile Phone Sensing Middleware: Learning from a Large-Scale Experiment," in *ACM/IFIP/USENIX Middleware 2016*, 2016.

[42] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[43] Z. Song, C. H. Liu, J. Wu, J. Ma, and W. Wang, "Qoi-aware multitask-oriented dynamic participant selection with budget constraints," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 9, pp. 4618–4632, 2014.

[44] D. Zhang, H. Xiong, L. Wang, and G. Chen, "CrowdRecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 703–714.

[45] H. Li, T. Li, and Y. Wang, "Dynamic participant recruitment of mobile crowd sensing for heterogeneous sensing tasks," in *2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 2015, pp. 136–144.

[46] Y. Liu, B. Guo, Y. Wang, W. Wu, Z. Yu, and D. Zhang, "TaskMe: Multi-task allocation in mobile crowd sensing," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 403–414.

[47] W. Zhu, W. Guo, Z. Yu, and H. Xiong, "Multitask allocation to heterogeneous participants in mobile crowd sensing," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.

[48] E. Luger and T. Rodden, "An informed view on consent for UbiComp," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ser. UbiComp '13, 2013, p. 529–538.

[49] A. Bechmann, "Non-Informed Consent Cultures: Privacy Policies and App Contracts on Facebook," *Journal of Media Business Studies*, vol. 11, pp. 21–38, 03 2015.

[50] B. Custers, S. Van der Hof, and B. Schermer, "Privacy expectations of social media users: The role of informed consent in privacy policies," *Policy & Internet*, vol. 6, 09 2014.

[51] E. Okoyomon, N. Samarin, P. Wijesekera, A. E. B. On, N. Vallina-Rodriguez, I. Reyes, Á. Feal, and S. Egelman, "On the ridiculousness of notice and consent: Contradictions in app privacy policies," in *Proc. Workshop on Technology and Consumer Protection (ConPro '19)*, 2019.