Privacy-Preserved Data Sharing Towards Multiple Parties in Industrial IoTs

Xu Zheng, *Member, IEEE*, and Zhipeng Cai, *Senior Member, IEEE*

***Abstract*— The effective physical data sharing has been facilitating the functionality of Industrial IoTs, which is believed to be one primary basis for Industry 4.0. These physical data, while providing pivotal information for multiple components of a production system, also bring in severe privacy issues for both workers and manufacturers, thus aggravating the challenges for data sharing. Current designs tend to simplify the behaviors of participants for better theoretical analysis, and they cannot properly handle the challenges in IIoTs where the behaviors are more complicated and correlated. Therefore, this paper proposes a privacy-preserved data sharing framework for IIoTs, where multiple competing data consumers exist in different stages of the system. The framework allows data contributors to share their contents upon requests. The uploaded contents will be perturbed to preserve the sensitive status of contributors. The differential privacy is adopted in the perturbation to guarantee the privacy preservation. Then the data collector will process and relay contents with subsequent data consumers. This data collector will gain both its own data utility and extra profits in data relay. Two algorithms are proposed for data sharing in different scenarios, based on whether the service provider will further process the contents to retain its exclusive utility. This work also provides for both algorithms a comprehensive consideration on privacy, data utility, bandwidth efficiency, payment, and rationality for data sharing. Finally, the evaluation on real-world datasets demonstrates the effectiveness of proposed methods, together with clues for data sharing towards Industry 4.0.**

***Index Terms*— Industrial IoTs, differential privacy, data sharing, Industry 4.0.**

I. Introduction

TO SUPPORT the emergence and development of next-generation industry, *i.e.*, Industry 4.0, one essential condition is the communication interface for seamlessly flow of information throughout the production system [1]. The involvement of IoTs has made significant facilitation to this trend, where the production data could be ubiquitously collected and shared. It brings together a novel eco system noted as Industrial Internet-of-Things (IIoTs) [2]. A typical IIoT system may involve categories of industrial data, ranging from the traffic records originated from the freight systems or regular crowds [3], to the sophisticated production data from the pipelines in factories and supplement chains. These data can facilitate intelligentization for numerous industrial decisions [4] including the smart logistics, scheduling and supplement of resources, adaption of investment, *etc*. However, the data also implicates private information and business profits for data owners, significantly hindering the seamless flow of information. Careless sharing strategies may disclose the sensitive status for workers, or providing extra information where subscribers can apply for maliciously commercial purpose. Therefore, this work studies the communication interface for privacy-preserved data sharing within industrial domains, considering the existence of parties in multi stages.

The data sharing has been a pivotal aspect for IIoTs. Generally, a typical IIoT includes some data generators, like the factories, the workers, or individual contributors. They actively contribute contents of different categories. For example, the scales of freight and traffic flows within different regions of a city. These contents can benefits multiple aspects of IIoTs, like the adaptive deployment of new logistic capabilities, the real-time scheduling of product transportation, the decision of locations for new factories, and also the strategies of production for upstream and downstream parties. However, the data may also correlate with the privacy issues, referring to the sensitive status of contributors. For example, it may link to the daily movement of a driver [5], the behavior patterns for people in a community, or the potential plans of some companies. These threats have been challenging the adoption of IIoTs.

Furthermore, the applications on IIoT data could also be more complicated [6] than basic IoT systems. The implementation of the new era industrial system requests a continuous chain of data sharing in different stages, starting from the material supplier all the way to the the final retailers. Therefore, data will usually be flown among multiple parties. For example, freight companies apply the traffic loads for service deployment and real-time scheduling, while the upstream and downstream subscribers may further apply the information to allocate their resource for production and retailing. In this case, the data is leveraged for multiple applications, thus involving both more utilities and the extra leakage of privacy in the data.

Unfortunately, current study for sensitive data sharing in IoTs mainly focuses on data perturbation and data trading. The perturbation mechanisms conceal the sensitive status like the locations or attributes. The trading mechanism usually tries to achieve a balanced payment [7], where the privacy in data

Manuscript received July 16, 2019; revised January 14, 2020; accepted February 26, 2020. Date of publication March 16, 2020; date of current version May 7, 2020. This work was supported in part by the National Science Foundation (NSF) under Grant 1912753, Grant 1741277, Grant 1829674, and Grant 1704287 and in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 61802050. *(Corresponding author: Zhipeng Cai.)*

Xu Zheng is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: xzheng@uestc.edu.cn).

Zhipeng Cai is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303 USA (e-mail: zcai@gsu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2020.2980802

could be evaluated and the objective is usually to maximize the social welfare. However, these works usually ignore the multiple applications of data, where service providers and subscribers all rely on the data while service providers also trade the data for extra profits.

Actually, there are two major challenges for data sharing in aforementioned IIoTs. The first one is to achieve the balance among the privacy, the profits, and the utility of data. Workers will not wish to yield their privacy too much upon unreasonable profits, while the service provider and subscribers try to maximize their benefits under constrained budgets, i.e., to retain the utility in the collected results. Furthermore, the service provider itself has to make a balance between the ratios of data and the accuracy of data collected from workers, to make the estimated results more applicable. Therefore, there is another inherent trade-off under limited budgets. Both trade-offs request a comprehensively designed scheme for data sharing. Secondly, during the data sharing between the service provider and subscribers, the service provider also expects to both benefit from the data relay and retain its trade secret toward subscribers, which could potentially be its rivals. The service provider also has to make sure the subscribers will not bypassing them and directly collect data from workers. In this circumstance, workers are higher paid, but they are considered as less trustworthy as no authorization are provided by the service provider. The subscribers are lower-charged as no extra cost are charged by the service provider, while also suffering relatively low utility on the unreliable data. Consequently, the service provider must carefully design the data sharing strategy to avoid this potential collusion. As far as we know, none existing results can properly cover both challenges.

To mitigate the gap, this paper propose a novel framework for data sharing in IIoTs. Specifically, the three-party model is applied, where workers, one service provider, subscribers are involved. The workers hold multiple contents in the system. The service provider collects contents from workers, retains them for own benefits, and re-trades the data to subscribers for extra profits. The subscribers request the data for their own production process, and pay for the data according to the charge of the service provider. This model could also be extended for more parties. Furthermore, the contents are considered as private information for each worker. In this paper, the differential privacy is adopted to preserve the sensitive information for workers. which allows the adversaries with strongest background knowledge. Specifically, the service provider and subscribers concern on the counting and the histogram distribution on underlying contents, which are essential and major requests for IIoTs as they can both provide a meaningful sketch for underlying systems.

Specifically, we assume the workers are paid according to the accuracy and ratio of uploaded contents, *i.e.*, the privacy and the bandwidth consumption. They will sample, perturb and upload the contents under local differential privacy, according to the received privacy budget and payments. Then the service provider calculates the results according to the collected data, and will further process and share the data with subscribers. The goal for the service provider is to first derive a proper strategy for data collection by achieving good data utility. This is studied in our first scenario, where the service provider and subscribers are tied in the data sharing. In the second scenario, the service provider will further perturb the result and relay it with subscriber. Then the goal also includes the retaining of exclusive profits for service provider, while workers and subscribers are fairly treated. In this case, the subscribers may bypass the service provider and collude with workers. The service provider guarantees that subscribers will not achieve better utility by collusion with workers. This paper propose two corresponding algorithms designed for two scenarios, each with comprehensive analysis on performance Finally, the framework are validated on real-world datasets. The results validate the effectiveness of proposed algorithms. As far as we know, this is the first study combing the privacy, utility, the fairness, the payment among multiple parties in IIoTs. The main contribution of this work includes:

- A novel framework for privacy-preserved data sharing is proposed for IIoTs.
- The problem of data sharing by joint consideration on the utility, privacy, and the rationality is formulated.
- Two algorithms are proposed for different scenarios of data sharing, covering major situations for IIoTs.
- Corresponding analysis is proposed to demonstrate the utility of data sharing.
- Extensive evaluation is conducted on real-world dataset, and the results reveal the effectiveness of the proposed methods.

The remainder of this paper is organized as follow. Section II overviews the literature on corresponding research topics. Section III gives our problem formulation, addressing two scenarios. Section IV provides the algorithm and analysis for the first scenario. Section V introduces the algorithm designed for the second scenario. Section VI provides the evaluation results. Finally, section VII concludes the whole paper.

II. Related Work

The data sharing has long been considering as a primary requirement for facilitating the functionality of IIoTs [8]. For example, it is believed that the freight and taxi flows could work as a reference for traffic loads and freight capabilities for urban area [9]–[11], and to continuously tracking packages in logistics system will be critical to learn the life cycle of a product [12]. Owing to these benefits, numerous studies are conducted to investigate the data sharing within IIoT systems. These works focus on the novel applications of shared data [13], the resource allocation during the data exchange [14], and the design of integrated system [15] where data can be freely shared. However, one common drawback of these studies is the ignorance of privacy issues, which is even more serious due to the involvement of agnostic physical data [16].

There has also been a long study history for privacy preservation in IoTs [17]. The location information, as an inherent attribute for many physical data, has been thoroughly studied to thwart attacks from different adversaries. Both the sensitive positions and the private knowledge beneath these locations are investigated and properly handled [18]. Meanwhile, other types of physical data has also been treated, and potential threats are suggested. For example, the gesture data could reveal the pin number for a mobile phone users [19], as well as reveal the application a user is currently focusing on. Meanwhile, the efficiency has usually be jointly considered with privacy preservation for IIoTs [20], [21]. However, this situation is even more urgent for IIoTs, as this kind of systems involves huge domains of physical data [22], [23]. and the correlations among multiple parties are more sophisticated, which makes existing solutions inappropriate for data sharing in IIoTs.

As for the privacy-preserved data sharing in distributed environment [24], LDP [25] is believed to achieve the optimal performance by allowing an arbitrary background knowledge for adversaries. Google proposed Basic RAPPOR [26] and RAPPOR [26], both of which provide LDP for participated users. These two methods are based on the idea of random response, and differ in their bandwidth efficiency. Subsequent studies apply the LDP for multiple purpose of data sharing. For example, the histogram distribution [27], the general graph structures [28], and the frequent items [29], [30]. There are even some works [31] trying to conclude current studies on LDP and providing guidelines for applications. Another body of studies try to involve the links among multiple parties [32] during the IoT data sharing. They either consider the privacy preservation [33] or the utility maintenance [34] as a common task among participants. However, the cases for data sharing in IIoTs is usually more complicated, where multiple parties may compete and hold diverse purposes for data.

Finally, the crowd-sensing has also been treated as a promising approach for data sharing in IIoTs, where workers are allowed to voluntarily participate in the systems for data sharing and rewarding. Current studies aim at deriving plans where data utility, fairness, and social welfare can be jointly optimized. As for the privacy concerns, both the location privacy [35] and the privacy within the uploaded data [36] have been studied for the crowd-sensing systems. However, there is still a gap between crowd-sensing systems and the scope of IIoTs, where the service provider may also make use of the collected contents, and compete with subscribers on markets. Therefore, algorithms and strategies designed for crowd-sensing systems will be inapplicable.

III. Problem Formulation

This section gives basic settings in the framework. It includes the definition on workers, service providers, and subscribers. The second part gives the description of different scenarios. The third part introduce preliminaries for differential privacy.

A. Basic Settings

There are three parties in the framework: the workers, the service provider, and the subscribers. The workers are also named as data contributors, while the service provider and the subscribers are also combined as data consumers.

1) Workers: The workers, denoted as $\{u_1, u_2, \cdots, u_M\}$, participate in the system and generate contents about different applications. For example, the traffic data uploaded by each drivers, or the category of daily manufacturing data by different workshops. Each worker u_i separately generates a set of contents $C_i = \{c_{i1}, c_{i2}, \cdots, c_{iK_i}\}$, which could be daily visited positions of a drivers, the types of manufacturing products, *etc*. All contents are selected from a content pool $F = \{f_1, f_2, \cdots, f_{K_0}\}$, which includes all distinct contents. The data will be transmitted in the whole system for further utility.

2) Attacking Model: The workers will concern on their private contents, which may correlate with their personal status. For example, the mobility patterns for drivers, or some specific types of products. In this work, the adversaries are malicious data viewers who try to infer the personal status of workers. They are both honest and curious, and can be either the service provider or subscribers. Our framework involves the idea of differential privacy, which achieve the state-of-the-art guarantee for privacy preservation. Generally, this metric allows a data owner to publish a perturbed statistic for her contents, without disclosing any one of her individual contents. Differential privacy assumes the strongest background knowledge for adversaries, *i.e.*, they are allowed to know all contents except for the target one, and none meaningful information will be disclosed from the statistics. Furthermore, this paper adopt the local differential privacy, where data can be collected in distributed manners. The formal definition of local differential privacy is shown as follow:

Definition 1 (Local Differential Privacy [25]): An algorithm or an encoding function Q satisfies ϵ-local differential privacy (ϵ-LDP) where $\epsilon \geq 0$, if and only if for any contents T_i and T_j, we have

$$\forall y \in Range(Q) : Pr[Q(T_i) = y] \leq e^{\epsilon} Pr[Q(T_j) = y],$$

where Range(Q) denotes the set of all possible outputs of the algorithm or encoding function Q.

In LDP, ϵ_i is the privacy budget or privacy factor for data sharing. Workers accept larger ϵ_i by admitting a loose preservation on their contents, and less noise will be added. Furthermore, we also introduce the parallel properties for differential privacy.

Theorem 1 (Parallel Composition [37]): Let $\{t_{i1}, t_{i2}, \cdots, t_{iK_i}\}$ be the contents held by user u_i, and f_is be a set of K_i encoding functions each providing ϵ_i-differential privacy. Then applying all f_is to their corresponding contents t_{ij} can guarantee a $\max\{\epsilon_i\}$-differential privacy for u_i.

This properties indicates that content from one worker will not affect the privacy for others when they are merged, as their contents are disjoint.

3) Profits: Finally, workers share their data for profits. The service provider requests the ratio of uploaded contents and the scales of injected noises for each worker. Workers process and upload contents for the rewards accordingly. The charge for each content is defined as $R(\epsilon_i)$, where a larger ϵ_i leads to larger charge since more privacy is disclosed.

4) Subscribers: The subscribers $S = \{S_1, S_2, \cdots\}$ act as the data consumers in the system. There could be multiple subscribers, and their correlations could be either identical or in sequential manners. In the later case, a subscriber could be both data consumers and data owners for subsequent subscribers [38]. In our framework, the subscribers expect to retrieve knowledge about the contents within the production system.

Due to the concern on privacy and profits, subscribers need to pay for the data. Subscribers will pay according to the accuracy of the data. Therefore, they will pay more to achieve better data utility. Assume the query to be $Q(\mathscr{C})$, where $\mathscr{C}$ is the combination of all data generated in the system, and $Q(\cdot)$ is the function deriving the necessary information for subscriber. For example, $Q(\cdot)$ could be the histogram or the counting of each distinct categories of contents. Subscribers will not receive the accurate results, due to the limited budget and the privacy concerns. Therefore, the utility of the retrieved result is determined by the expectation and the variance, where

$$E(Q(\mathscr{C}') - Q(\mathscr{C})) = 0, \tag{1}$$

meaning the results are unbiased, and $Var(Q(\mathscr{C}')$ should be minimized.

5) Service Provider: To maintain the functionality of the data flow for IIoTs, one service provider should be adopted to coordinate the data sharing among workers and subscribers. The service provider could be some dominating third parties like the power companies within the supply chain. The service provider will request the data from workers, and trade it with other subscribers. Meanwhile, the service provider will also take the data for their own profits, as they are also involved in the IIoTs. Specifically, assume the total budgets received from subscribers to be B_0. Then the service provider will directly apply the budget to collect a proportion of contents from workers, with some degree of privacy constraints. The received contents will be processed to derive the necessary knowledge, denoted as $Q_{sp}(\mathscr{C}')$. Finally, the service provider will forward the results to subscribers, or further process the results before forwarding, to retain their own profits against potential rivals.

B. Scenarios

Two major scenarios of data sharing are considered for IIoTs. The first one is the coalitional model, where the service provider and following subscribers will be tied and share identical information. This is the case where the service provider is just a basic platform for data fusion in IIoTs [39], and will not participate in the functionality of the production system. This scenario also holds when the service provider serves different departments of a large enterprise. The second scenario is the independent model, where the service provider and subscribers are independent parties and potential rivals. The service provider processes the received results before sharing it with subscribers.

1) Coalitional Model: This scenario assumes there is one service provider and one subscriber to simplify the underlying data flow. The two parties are tied with each other, such that they will share exactly the same information. The service provider SP_0 will first receive the budget B_0 from the subscriber S_0. The budget will be paid to different workers, and these workers will upload their contents to the service providers. Finally, the service provider processes the received contents and shares the results with subscribers.

The service provider needs to decide the strategies for data collection among workers. Specifically, two parameters should be considered: the acquisition ratio β_0 and the privacy factor ϵ_0. The acquisition ratio β_0 indicates the scales of contents uploaded to SP_0, and the privacy factor ϵ_0 indicates the degree of differential privacy adopted by each worker. Generally, a larger β_0 means more workers will participate in the data sharing, while a larger ϵ_0 means workers will provide more accurate contents. However, the payment $R(\cdot)$ for each worker is also determined by ϵ_0, where a larger ϵ_0 leads to higher cost. Therefore, the service provider has to decide the ratio of contents and the payment for corresponding workers under the limited budget.

The service provider and the subscriber try to extract the counting of each categories of contents exists in the whole system. This kind of information will provide overview of the situation in the system, like the distribution of trucks in different regions, or the types of products for manufacturing factories. Therefore, we have $Q(\cdot) = \{Co_1, Co_2, \cdots, Co_{K_0}\}$. Assume the original counting for each types of contents as $\{Co_1^0, Co_2^0, \cdots, Co_{K_0}^0\}$. Then the utility of the service provider and subscriber is defined in equation (2).

$$\frac{\sum_{i=1}^{K_0} |\frac{Co_i - Co_i^0}{Co_i^0}|}{K_0}, \tag{2}$$

which is the mean of relative errors on all categories.

As the data collection is one snapshot query, the distribution of the derived results will be applied to evaluate the effectiveness of the results. Therefore, our objective is to determine the set of acquisition ratio β_0 and privacy factor ϵ_0, by given a fixed budget B_0, such that the strategy can minimize the variance for the results.

2) Independent Model: The second scenario assumes that the service provider and the subscriber are independent parties. The subscriber will first propose its budget to the service provider, which will apply the budget to acquire contents from workers and extract corresponding results. Then this service provider further process the results and conceal some private business information, before sharing it with the subscriber. To simplify and clarify the analysis, we still consider the case with one service provider S_0 and one subscriber SP_0.

In this scenario, the service provider SP_0 still collects the counting for different categories of contents, and estimates the utility of the results according to the accuracy. Meanwhile, the subscriber S_0 will request for the histogram distribution of the counting $Q_S(\cdot)$, which will also provide some clues for the system. Therefore, SP_0 will further process the results for counting query, and share the processed histogram graph to S_0. To be general, our framework assumes SP_0 concerns on the counting for each category, and treats it as the private information. The total utility of SP_0 is $U_{SP} = U_{content} + U_{relay}$, where $U_{content}$ refers to the accuracy of results, and U_{relay} refers to the retained information towards S_0, which is determined by the scale of noise injected in the results.

A second objective for SP_0 is to also ensure a reasonable benefits for the subscriber S_0, where S_0 cannot bypass SP_0 and gain better results by directly collecting data from the workers. Assume the payment for selling data to SP_0 as $R(\cdot)$, the charge for SP_0 from S_0 as $R_{SP}(\cdot)$, and the payment for selling data to S_0 as $R_S(\cdot)$. For a fixed ϵ_0,

$$R(\cdot) \leq R_S(\cdot) \leq R_{SP}(\cdot), \tag{3}$$

which follows the general rules in real-world data trading. Meanwhile, the subscriber S_0 possibly suffers the low data utility when directly requesting data from workers, owing to the absence of a trustable data platform SP_0. Our framework formulates the general data utility factor as γ_0, meaning the data directly requested from workers could amplify the expected variance by γ_0.

Generally, in the second scenario, the objective is as follow: SP_0 applies the budgets from S_0 to collect the data from workers, *i.e.*, determines the acquisition ratio β_0 and privacy factor ϵ_0. Then SP_0 adopts another privacy factor ϵ_0' for result perturbation before exchanging the results with S_0. The corresponding strategies guarantee the subscriber S_0 cannot bypass SP_0 for better utility, while the utility for SP_0 is maximized, *i.e.*, a minimum ϵ_0'.

C. Preliminaries

This part addresses the randomized response mechanism, which is a typical method for data collection under LDP.

Assume there is a L-bits vector with binary entry, denoted as $V = (v_1, v_2, \cdots, v_L)$.

Then V' can be generated by randomized response:

$$Pr[V'[i] = 1] = \begin{cases} 1 - \dfrac{1}{2}f, & if\ V[i] = 1 \\ \dfrac{1}{2}f, & if\ V[i] = 0. \end{cases} \tag{4}$$

Actually, this mechanism of perturbation achieves LDP property for vector V, which is proved by previous work [31]:

Theorem 2: For an arbitrary vector $V = (v_1, v_2, \cdots, v_L)$, the randomized response achieves ϵ-LDP for $\epsilon = ln((\frac{1-\frac{1}{2}f}{\frac{1}{2}f})^2)$.

IV. Solutions for Utility-Optimized Data Collection in IIoTs

This section provides the algorithm for data collection in the first scenario, where the service provider and the subscribers are tied in the data sharing. It first introduces the details of the proposed algorithm, and proposes corresponding analysis demonstrating the effectiveness of the algorithm.

A. Algorithm for Data Collection

In the first scenario, the service provider and the subscriber are jointly considered for one objective. Therefore, this part applies SP_0 to indicate both sides for simplicity. Initially, SP_0 collects contents from M individual workers $\{u_1, u_2, \cdots, u_M\}$ with a total budget B_0. The overview of the procedure is as follow. SP_0 first determines the ratio and the set of contents to be upload. Then workers with selected contents will receive the payment and corresponding requirements ϵ_0 on privacy. These selected workers upload the perturbed contents to SP_0. Finally, SP_0 aggregates the contents and estimates the final results. To be clear, the algorithm is denoted as the ***Joint Optimized Data Collection*** algorithm (JODC for short).

In the first step, SP_0 derives the acquisition ratio β_0 and privacy factor ϵ_0. Upon the budget B_0, the cost function $R(\cdot)$, and the scale of content $\sum K_i$, SP_0 determines the corresponding β_0 and ϵ_0 such that

$$\{\beta_0, \epsilon_0\} = \underset{\beta,\epsilon}{\operatorname{argmin}}\, Var(\sum_{i=1}^{K_0} Co_i), \tag{5}$$

where the derived results achieve minimum variance for the estimation.

In the second step, SP_0 will randomly pick $\sum K_i \cdot \beta_0$ contents without knowing the details of them, and the owner of these contents are selected as the data contributors. SP_0 distributes the request containing ϵ_0, together with the payment to these workers. In JODC, SP_0 will iteratively sample one content and add the worker into the candidate set, until $\sum K_i \cdot \beta_0$ distinct contents are selected.

In the third step, the selected workers in the candidate set receive the privacy factor ϵ_0. Each selected worker u_i will first locally process their contents. Each content will be perturbed with random response introduced in formula (4). Therefore, the selected content set $C_i = \{c_{i1}, c_{i2}, \cdots, c_{iK_i'}\}$ will be obfuscated to $C_i' = \{c_{i1}', c_{i2}', \cdots, c_{iK_i'}'\}$ and uploaded to SP_0.

Finally, SP_0 will sum up each category of contents within the collected data. Assume the accumulated numbers to be $\{Co_1', Co_2', \cdots, Co_{K_0}'\}$. SP_0 will scale up each counting by

$$Co_i = \frac{Co_i' - N \cdot \beta_0 \cdot 1/2f}{(1-f) \cdot \beta_0}, \tag{6}$$

where N is the total number of contents in the system. Finally, SP_0 gets the outputs

$$\{Co_1, Co_2, \cdots, Co_{K_0}\}. \tag{7}$$

The pseudo code for JODC is shown in algorithm 1.

B. Analysis

This subsection first analyzes the temporal, spatial, and bandwidth efficiency of JODC. Then the privacy preservation for each worker is demonstrated. The third part discusses the accuracy for the derived results, and the last part analyzes the effectiveness of the extracted results for the service provider.

1) Efficiency: The time for the first step is $O(1)$, where SP_0 extracts the necessary parameters. In the second phase, the time for selecting contents is $O(N \cdot \beta_0) = O(N)$, as SP_0 takes $N \cdot \beta_0$ rounds to pick up contents. In the third phase, each worker u_i consumes $O(K_i)$ to perturb all her contents. Therefore, the time complexity for the third phase it $O(\max K_i)$. Finally, the last phase requests SP_0 to scale up the counting for all categories, which is $O(K_0)$. Generally, the time complexity for JODC is $O(N + K_0)$.

Algorithm 1 Joint Optimized Data Collection

```
$SP_0$ derives the sampling ratio $\beta_0$ and privacy budget $\epsilon_0$ according to equation (5).
for Each $u_i$ do
    Sort contents in $C_i$ in an arbitrary order.
    Upload $K_i$ to $SP_0$.
end for
$SP_0$ arranges a unique number within $[1, \sum K_i]$ to each content.
$SP_0$ randomly picks $\sum K_i \cdot \beta_0$ samples.
$SP_0$ distributes numbers of selected contents to $u_i$s.
for Each $u_i$ do
    for Each selected content $c_{ij}$ do
        Set $f$ according to Theorem 2.
        Execute random response on $c_{ij}$ with $f$.
    end for
    Upload $C'_i$ to $SP_0$.
end for
for Each $Co_i$ do
    $SP_0$ estimates $Co_i$ according to equation (6)
end for
```

The spatial complexity for SP_0 is $O(M \cdot N)$, as it needs to store the contents for all selected workers. However, JODC can further apply the incremental strategy to reduce the spatial complexity to $O(K_0)$, where the counting is added upon each received content set. The spatial complexity for each worker is $O(N)$, where the worker needs to keep a record for each of her contents.

Finally, the total bandwidth consumption for JODC is $O(M + M \cdot N \cdot K_0)$, where SP_0 distribute the requests to $M \cdot \beta_0$ workers, and each worker consumes $O(NK_0)$ bandwidth to return the content set.

2) Privacy Preservation: As JODC follows the typical random response mechanism, it can preserve each worker under differential privacy within the uploaded contents. Furthermore, the sequential property guarantees the aggregated results will reveal no extra information for workers. Therefore, JODC can preserve each worker under the following theorem.

Theorem 3: Within JODC, each worker will be preserved under ϵ_0-differential privacy, where ϵ_0 is the privacy factor set by SP_0.

3) Accuracy: The accuracy of the results mainly comes in two folds. First, the estimated results should be an unbiased estimator, *i.e.*, the service provider requests the expectation of the results to be an unbiased estimation for the ground truth:

$$E(Co_i) = |Co_i^0|, \quad \forall i \leq K_0 \tag{8}$$

where $|Co_i^0|$ indicates the true number of contents belonging to the category Co_i. Second, the service providers hope to extract the results with strong stability, which means the results should be likely distributed around the ground truth. As for the second property, the variance will be adopted to show the stability.

As for SP_0, the uncertainty of results are introduced by two steps: the content selection, and the content perturbation. Actually, both steps follow the basic idea of Bernoulli sampling, and they are independent variables. The following theorem guarantees the unbiased results for SP_0.

Theorem 4: For each category of content Co_i, JODC can provide an unbiased estimation under given β_0 and ϵ_0.

Proof: Initially, the estimation for each category of content follow the equation below.

$$Co_i = \frac{Co'_i - N \cdot \beta_0 \cdot 1/2f}{(1-f) \cdot \beta_0}, \tag{9}$$

where Co'_i indicates the total number of contents collected from workers, which belong to category f_i, *i.e.*, the total number of vectors with ith entry equals 1.

We further denote N_i as the total size of contents belonging to category f_i in the system. According to the definition fo random response, we have

$$Co'_i = N_i \cdot V_s \cdot V_r + (N - N_i) \cdot V_s \cdot V'_r, \tag{10}$$

where V_s is the sampling variable for whether a content is selected, V_r and V'_r indicates whether the random response will retain or reverse the corresponding bit standing for the content in f_i.

Now we define the following symbol:

$$\Phi = Co_i - N_i. \tag{11}$$

Then we have

$$\begin{aligned} E(\Phi) &= E(Co_i - N_i) \\ &= E(\frac{Co'_i - N \cdot \beta_0 \cdot 1/2\ f}{(1-f) \cdot \beta_0}) - N_i \\ &= \frac{E(Co'_i) - N \cdot \beta_0 \cdot 1/2\ f}{(1-f) \cdot \beta_0} - N_i \\ &= \frac{E(N_i V_s V_r + (N - N_i) V_s V'_r) - N\beta_0 1/2\ f}{(1-f) \cdot \beta_0} - N_i, \end{aligned}$$

As V_s, V_r and V'_r are independent variables,

$$\begin{aligned} &E(N_i V_s V_r + (N - N_i) V_s V'_r) \\ &\quad = N_i E(V_s) E(V_r) + (N - N_i) E(V_s) E(V'_r) \\ &\quad = N_i \beta_0 (1 - \frac{1}{2} f) + (N - N_i) \beta_0 \frac{1}{2} f. \end{aligned}$$

Therefore,

$$E(\Phi) = \frac{N_i \beta_0 (1-f)}{(1-f) \cdot \beta_0} - N_i = 0. \tag{12}$$

Then we have

$$E(Co_i) = N_i,$$

which means Co_i is an unbiased estimator. ■

According to Theorem 4, JODC provides an unbiased estimation for each category of contents. Therefore, the final outputs $\{Co_1, Co_2, \cdots, Co_{K_0}\}$ will also be an unbiased estimator, as the sampling on different contents is independent.

The variance of the estimated results are calculated in lemma 1. The main idea of the lemma is to combine the variance from two steps of sampling, and derive the correlation between the variance and the two parameters β_0 and ϵ_0.

Lemma 1: For each kind of contents Co_i, with parameters β_0 and ϵ_0, the variance follows $Var(Co_i) \leq \frac{N_i^2+N^2\frac{1}{2}f}{(1-f)^2\beta_0} + \frac{1}{4}\frac{Nf^2}{(1-f)^2} + \frac{2NN_if}{(1-f)^2}$

Proof:

First of all, we have

$$Var(\Phi) = Var(Co_i - N_i) = Var(Co_i), \tag{13}$$

as N_i is a constant. Since $E(\Phi)^2 = 0$, we have

$$\begin{aligned} Var(\Phi) &= E(\Phi^2) - E(\Phi)^2 = E(\Phi^2) \\ &= E[(\frac{Co_i' - N\beta_0 1/2\ f}{(1-f)\cdot\beta_0} - N_i)^2] \\ &= E[(\frac{Co_i' - N\beta_0 1/2\ f}{(1-f)\cdot\beta_0} - N_i)^2] \\ &= N_i^2 + E(\frac{(Co_i' - N\beta_0 1/2\ f)^2}{(1-f)^2\cdot\beta_0^2}) \\ &\quad - 2N_i \cdot E(\frac{Co_i' - N\beta_0 1/2\ f}{(1-f)\cdot\beta_0}). \end{aligned}$$

As V_s is a Bernoulli samplings,

$$E(V_s^2) = E(V_s). \tag{14}$$

The same conclusion also holds for V_r and V_r'.

Therefore,

$$\begin{aligned} &Var(\Phi) \\ &= N_i^2 + \frac{E(Co_i' - N\beta_0 1/2\ f)^2}{(1-f)^2\cdot\beta_0^2} - 2N_i \cdot N_i. \\ &= \frac{E(Co_i'^2) - N^2\beta_0^2 1/4\ f^2 - N\beta_0\ fE(Co_i'))}{(1-f)^2\cdot\beta_0^2} - N_i^2 \\ &= \frac{N\beta_0\ f}{(1-f)^2\cdot\beta_0^2}(N_i\beta_0(1-\frac{1}{2}f) + (N-N_i)\beta_0(\frac{1}{2}f)) \\ &\quad - N_i^2 - \frac{N^2\beta_0^2 1/4\ f^2}{(1-f)^2\cdot\beta_0^2} + \frac{N_i^2\beta_0(1-\frac{1}{2}f)+(N-N_i)^2\beta_0\frac{1}{2}f}{(1-f)^2\cdot\beta_0^2} \\ &\quad + \frac{N_i(N-N_i)\beta_0^2(1-\frac{1}{2}f)f}{(1-f)^2\cdot\beta_0^2} \\ &\leq \frac{N_i^2 + N^2\frac{1}{2}f}{(1-f)^2\beta_0} + \frac{1}{4}\frac{Nf^2}{(1-f)^2} + \frac{2NN_if}{(1-f)^2} \end{aligned}$$

■

Finally, the following theorem demonstrates the total variance for the final outputs, which can be directly derived by accumulating the variance as they are independently determined.

Theorem 5: totalvar) The total variance for all categories of contents in the final output is less than $\frac{\sum N_i^2+K_0N^2\frac{1}{2}f}{(1-f)^2\beta_0} + \frac{K_0Nf^2+8N^2f}{4(1-f)^2}$

4) Effectiveness: Finally, the service provider expects high utility for the final output, which makes the algorithm more effective. Generally, this requirement could be achieved by reducing the total variance. This is same with the following problem:

Problem 1: Given a fixed budgets, how to derive the corresponding acquisition ratio β_0 and privacy factor ϵ_0, such that the total variance could be minimized.

With all the information available, JODC could iteratively derive the optimal results via the comparison on the combinations of different parameters. Intuitively, larger β_0 and ϵ_0 lead to better utility, as well as higher cost. Therefore, JODC may tune the parameters under the budget constraint, and the optimal performance can be achieved as the variance is a monotonic function for both parameters.

However, the design of optimal factors are in fact non-trivial as the variance is also correlated with the ground truth for each category, which is exactly a dilemma for the service providers. To mitigate this gap, one potential solution is to apply an approximation for the total variance, and then estimate the parameters accordingly. For example,

$$Var(\sum Co_i) \leq \frac{(K_0+1)N^2\frac{1}{2}f}{(1-f)^2\beta_0} + \frac{K_0Nf^2 + 8N^2f}{4(1-f)^2}, \tag{15}$$

where all information could be available for SP_0, and JODC solves the following problems:

Problem 2:

$$\min \ \frac{(K_0+1)N^2\frac{1}{2}f}{(1-f)^2\beta_0} + \frac{K_0Nf^2 + 8N^2f}{4(1-f)^2} \tag{16}$$

$$\text{s.t. } R(\epsilon_0)\cdot\beta_0\cdot N \leq B_0. \tag{17}$$

This problem could be solved by multiple methodologies like the Lagrange multiplier [40].

V. Solutions for Multi-Party Data Sharing in IIoTs

This section provides the algorithm for data collection in the second scenario, where the service provider and the subscribers are independent in the data sharing. It also starts with the details of the proposed algorithm, and proposes corresponding analysis demonstrating the effectiveness of the algorithm.

A. Algorithm for Multi-Party Data Sharing

In the second scenario, the service provider will first collect contents from workers, and relays the processed results to the subscriber. The general input for SP_0 is same with the first scenario. The output includes the derived counting for distinct contents for SP_0 and the histogram distribution for S_0. The overview of the algorithm is as follow. Initially, SP_0 receives the budgets B_0 from S_0. Then the service provider will follow the steps of JODC, and derive the results for itself. According to this results, SP_0 estimates the scales of information it can retained, and further processes the results to derive the final outputs, which will be forward to the subscriber S_0. The proposed algorithm is denoted as the ***Independent Optimized Data Collection*** algorithm (IODC for short).

In the first part, IODC processes exactly same with the JODC, where the budgets B_0 are distributed to different workers, and SP_0 gathers the contents to estimate the counting for different contents.

In the second part, SP_0 processes and relays the results to subscriber S_0.

The service provider first considers the case where S_0 directly collect data from workers. In this case, S_0 can retrieve contents with lower cost, which is $R_S(\epsilon)$, and $R(\epsilon) \leq R_S(\epsilon) \leq R_{SP}(\epsilon)$. However, S_0 also suffers the unreliability of contents, which is formulated by the amplifier γ_0 on the variance. Generally, SP_0 will apply the analysis for JODC to draw clues on the performance of direct data collection by S_0, where $R_S(\cdot)$ and γ_0 are considered.

Based on the derived results, SP_0 further estimates the gap between the current results and the optimal results for subscribers, which is shown in equation (18)

$$\begin{aligned} G &= Var_s(\sum C0_i) - Var(\sum C0_i) \\ &= (\gamma_0 - 1)Var(\sum C0_i), \end{aligned} \tag{18}$$

where $Var_s(\sum C0_i)$ is the variance when S_0 performs direct data collection. Then SP_0 further perturbs the histogram distribution according to the gap. The perturbation is implemented by a second random response on all its data with parameter ϵ_0', where the variance brought by ϵ_0' in the random response is $(\gamma_0 - 1)$ times of the original one.

Finally, SP_0 estimates the histogram based on the perturbed contents, where the same strategy in JODC will be applied to regulate the errors in random response. Then the results are forwarded to S_0.

B. Analysis

This subsection first analyzes the temporal, spatial, and bandwidth efficiency of IODC. The second part discusses the accuracy and rationality for the derived results. Finally, the privacy preservation for each worker is demonstrated.

1) Efficiency: This part focuses on the efficiency for the second phase of IODC, as the first part are similar with JODC algorithm.

In the second phase, it takes $O(1)$ to calculate the budget ϵ_0', and takes $O(NK_0)$ to perform the second random response. Therefore, the time complexity for the second phase of IODC is $O(NK_0)$.

Meanwhile, the spatial complexity of IODC is same with JODC, and it takes $O(K_0)$ bandwidth for SP_0 to release the results to S_0.

2) Accuracy and Rationality: This part studies the accuracy for SP_0 and S_0, and the rationality for S_0.

The service provider SP_0 follows the same strategies in JODC. Therefore, the accuracy for the results and the utility can be directly derived from the analysis in previous section. However, the accuracy for S_0 is more complicated as the results will be further processed before releasing to SP_0. Theorem 6 indicates that S_0 also receives unbiased results.

Theorem 6: Assume Co_i^S to be the estimated number of contents in category f_i for S_0, $E(Co_i^S) = N_i$.

Proof: The proof is straightforward. As both random response mechanisms introduce correction steps to regulate the extra contents in f_i, and apply a scaling factor to recall the flushed contents, they both provide unbiased estimation for the counting. Furthermore, as two mechanisms are independent, the final results are also unbiased. ■

Besides, we can also evaluate the variance for S_0, which is also a linear combination of variances introduced by two mechanisms.

Finally, we briefly discuss the rationality for the subscriber S_0. Intuitively, subscriber S_0 will get a lower cost and higher risky when directly requesting contents from workers. The service provider, as a trusted and trackable platform, can act as a connector for content sharing. It collects the contents from workers with a lower price, and exchange with S_0 on a relatively high price. However, as SP_0 also acts as the subscriber in the system, it will not wish to share exactly the same information with S_0 due to the concerns on profits. Therefore, further process will be adopted on the received contents before they are delivered to S_0. Theorem 7 shows the generated scheme for data sharing will be followed by subscribers.

Theorem 7: Under the processing of IODC algorithm, the subscriber S_0 cannot violate the framework with SP_0, and achieve better utility by directly requesting contents from workers.

Proof: Specifically, IODC is carefully designed such that the subscriber cannot achieve better utilities by bypassing SP_0. In IODC, SP_0 will first evaluate the performance when the bypassing occurs, and the variance of results for S_0 is evaluated based on the pricing function $R_S(\cdot)$ and the unreliability function γ_0. Then SP_0 obfuscates the results of the histogram distribution, such that the perturbed mechanism can guarantee the total variance from two phases will not be larger than the directly requesting. This property is achieved by equation (18) Therefore, both workers and S_0 will follow the framework with SP_0 in this case, and the rationality is achieved for S_0. ■

3) Privacy Preservation: The contents for each workers are only requested once, with privacy factor ϵ_0. The second random response are conducted purely on the collected contents, which will lead no more disclosure of privacy. Therefore, the following principle could be guaranteed.

Theorem 8: Under the processing of IODC algorithm, the contents for each worker is preserved with ϵ_0-differential privacy.

VI. Evaluation

This section introduces the evaluation results for the proposed methods. It starts with the introduction on adopted dataset, the compared methods, and the metrics. Then it gives the basic performance and the detailed evaluation for different scenarios.

A. Dataset and Settings

1) FAF4 Freight Datasets: The original data includes the transactions for freight happened within and outside United State, and is released on October 31, 2015 as a fourth version [41]. Each transaction includes the mode, the origin state, and the destination state of the transportation. This dataset may provide knowledge for the distribution and capabilities of nationwide logistic systems. Specifically, we extract transactions carried domestically from the whole dataset, and

TABLE I

CENSUS REGIONS AND DIVISIONS OF UNITED STATES

ID	Region Name	States
1	New England	CT, ME, MA, NH, FI, VT
2	Middle Atlantic	NJ, NY, PA
3	East North Central	IN, IL, MI, OH, WI
4	West North Central	ND, SD, NE, KS, MN, IA, MO
5	South Atlantic	DE, MD, DC, WV, VA, NC, SC, GA, FL
6	East South Central	AL, KY, TN, MS
7	West South Central	AR, LA, OK, TX
8	Mountain	AZ, CO, ID, MT, NM, NV, UT, WY
9	Pacific	AK, CA, HI, OR, WA

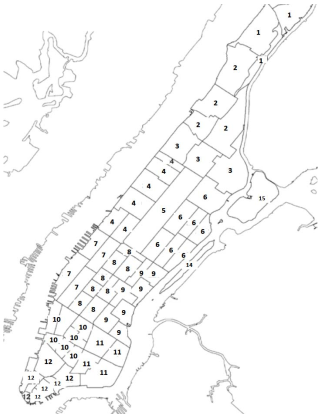
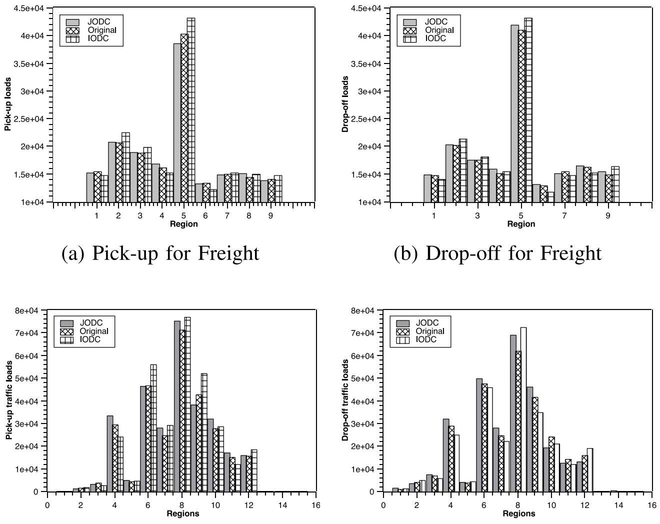


Fig. 1. Region partition.

constitute the freight flow datasets. The extracted dataset includes around 170K transactions. The census regions and divisions of United States are applied to partition the data into different categories. Table I shows the number and name of each category.

2) NYC Taxi Datasets: We also conduct the evaluation on the taxi traces in New York in the year of 2017 [42]. The taxi data could be treated as a strong reference for the urban traffic loads, thus being applied for numerous application like planning real-time routes for goods transportation, deploying extra capabilities for logistic systems, and selecting positions for new factories or storages. Specifically, the evaluation is conducted on daily traffic flows, and the regions are selected within the Manhattan district. We further partition the whole region into 15 sub-regions, each including a set of adjacent blocks. The partition of urban region is shown in Figure 1. The extracted dataset includes around 270K transactions.

3) Settings: The evaluation involves the scales of transactions started and ended in each state or sub-region, where the state and sub-region are the categories for transactions. In the evaluation results, JODC refers to the results received by service providers, and IODC refers the results received by subscribers as further process is conducted by service providers. To validate the effectiveness of the proposed algorithms, we compares the performance of both algorithms under various combinations of parameters. The involved settings majorly include the ratio of collected contents β_0, the privacy factor ϵ_0, and the perturbed factor ϵ_0'. Knowing that our algorithms apply the basic random response for privacy preservation, they can be easily extended for other sophisticated methods to achieve reduced variance.



(a) Pick-up for Freight

(b) Drop-off for Freight

(c) Pick-up for Taxi

(d) Drop-off for Taxi

Fig. 2. Observed traffic scales in different regions.

To measure the performance of each method, the relative errors between the observed traffic flows and the ground truths is applied as the metric. We further define the accuracy as 1 minus the the relative errors, to emphasize the positive performance. Each group of evaluation has been repeated twenty times to mitigate the influence from randomness. The average performance of accuracy for twenty rounds is used to indicate the stability of algorithms, and the accuracy for 20-round average traffic scale in each category is used to indicate the unbiased results of algorithms.

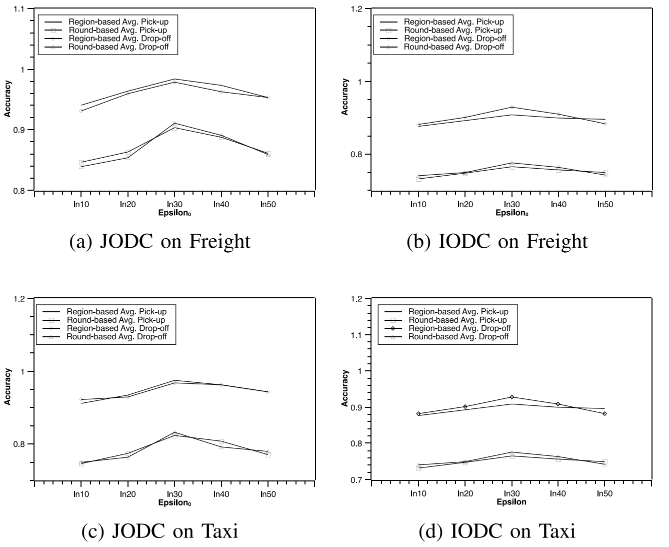
B. Basic Performance

This part shows the basic information provided for service providers and subscribers within both scenarios. The results can reveal whether the data subscribers can gain useful knowledge from the IIoT systems for further decision making. The privacy factor is set as $\epsilon = \ln 30$, the collection ratio is $\beta_0 = 0.3$, and the extra perturbation executed by the service provider in second scenario is $\epsilon' = 10$. Each algorithm has been run once and the results are shown in Figure 2.

According to the results, we can see that the observed results achieve a high consistency with the ground truth in both datasets. The service provider receives results with approximately 90% accuracy for the freight flow dataset (80% for the taxi flow dataset). The subscriber, even if its results is perturbed a second time by the service provider, still achieves around 85% (75% for the taxi flow dataset) accuracy on the results. We find the taxi flow suffers a slightly larger variance, as there are some minor regions with very small number of traces, and their errors will be more significant. Meanwhile, we also observe that regions with huge traffics receives higher accuracy. The underlying reason is that the impact of randomness will be alleviated as the number of contents increase.

C. Comparison on Different Parameters

This part investigates the impact of different combinations of β_0 and ϵ_0. Specifically, it studies the performance of



(a) JODC on Freight

(b) IODC on Freight

(c) JODC on Taxi

(d) IODC on Taxi

Fig. 3. Accuracy under various ϵ_0 and β_0.

different parameter combinations when the total budget is fixed. The budget for is set as $B_0 = \epsilon_0 \cdot \beta_0 \leq 1$. The evaluation sets $\epsilon_0 = \ln 10, \ln 20, \ln 30, \ln 40, \ln 50$, respectively. The results are shown in Figure 3.

As we see, the average accuracy for both cases first increases as ϵ_0 increases, indicating the scales of samples and the injected noise for each sample are well-balanced. Thus, the performance will be improved by approximately 5% to 10%. However, as ϵ_0 keeps increasing, the size of samples will be reduced, and leads to the reduction on performance. Consequently, the service provider needs to make a proper balance on the two factors, as is demonstrated in our analysis. Furthermore, we also find that the subscriber receives a little bit lower utility in IODC, due to the extra process by the service provider. However, the difference is insignificant as the adopted parameter ϵ_0', which is a relatively loose constraint.

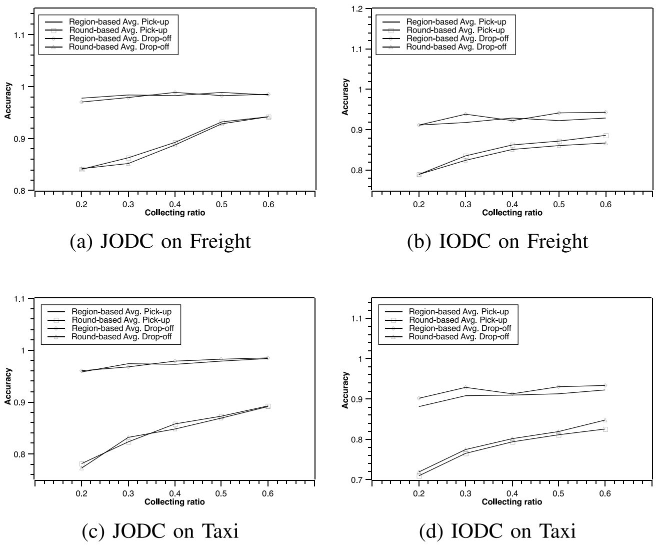
Finally, we also find that the accuracy is approaching 1 for both datasets, when we average the size of each category of contents among twenty repeats (*i.e.*, Region-based Avg.). This observation reveals that both of our algorithms could actually provide an unbiased estimation for the results. The barrier here is that to repeatedly request contents from workers will introduce both extra budgets and privacy issues.

D. The Impact of Samples and Privacy Concerns

This part studies the trend of the performance, as the ratio of collected contents and the privacy budget increase. The major objective for this part is to investigate whether it is worthy to devote more budgets for better utilities.

In this first group, the privacy factor is fixed as $\epsilon_0 = \ln 30$, and the ratio of collected contents range from 0.2 to 0.6, and incremented by 0.1. The results are shown in Figure 4.

As we see, the performance can be improved for both freight and taxi flows when the sampling size is increased. However, the speed for the improvement suffers a reduction, which means simply increasing the ratio of contents will not be a wise strategy for data collection. This is especially meaningful when the accuracy is already acceptable and the sampling ratio is relatively small. For example, the accuracy is approaching 85% for service providers and 80% for subscribers in taxi flows when sampling ratio is 0.4, as is shown in Figure 4(c) and Figure 4(d). However, the accuracy will be increased by less than 5 percents when the sampling size is increased to 0.5.



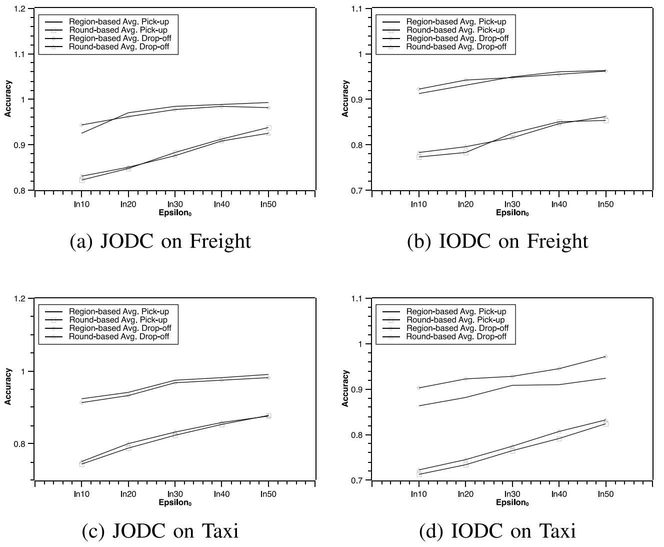
(a) JODC on Freight

(b) IODC on Freight

(c) JODC on Taxi

(d) IODC on Taxi

Fig. 4. Accuracy under various β_0.



(a) JODC on Freight

(b) IODC on Freight

(c) JODC on Taxi

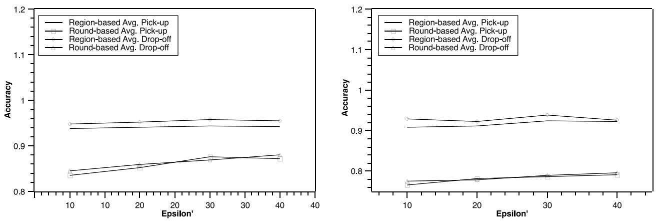
(d) IODC on Taxi

Fig. 5. Accuracy under various ϵ_0.

In the second group, the size of collected contents is fixed to $\beta_0 = 0.3$, and the privacy factors range from $\ln 10$ to $\ln 50$. The results are shown in Figure 5. As we see, similar conclusions can be drawn from the results. The effectiveness of applying extra budgets for more reliable contents will be insignificant. Generally, there will be a threshold where extra budgets will not bring much more benefits.

E. Impact of Profits for Service Providers

This part studies the utilities for subscribers when the service provider holds different opinions on its profits.



(a) Freight

(b) Taxi

Fig. 6. Accuracy under various ϵ_0'.

Specifically, the parameters are $\epsilon_0 = \ln 30$, and $\beta_0 = 0.3$. The scale of noise ϵ_0' introduced by the service provider changes from 10 to 40, indicating the service provider are more flexible. The performance is given in Figure 6. As is shown in the results, subscribers receive better utilities as ϵ_0' increases for both datasets, around 5% for freight flows and 3% for taxi flows. However, the improvement is still insignificant. One potential reason is that the service provider in this evaluation is already flexible on their profits, and will not inject heavy noise into the outputs.

VII. Conclusion

This paper investigates the problem of data sharing for IIoT systems, where multiple data consumers exist in different stages. These consumers request knowledge on the underlying system for different utilities. One consumer, acting as the service provider, will request data from workers, and relay the contents with other subscribers. Meanwhile, the privacy for workers are also considered, and they will perturb their contents before uploading. The perturbation mechanism is proved to provide differential privacy for workers. Two algorithms are designed for the data sharing, depending on the correlations between service providers and the subscribers. Both algorithms achieve a balance among the data utility, the payment, the privacy preservation, and the bandwidth consumption. The rationality for subscribers is also considered in the later case, where subscribers can rely on the service provider to achieve optimal data utility. Finally, the evaluation on real-world datasets has revealed the effectiveness of the proposed methods.

References

[1] Y. Lu, “Industry 4.0: A survey on technologies, applications and open research issues,” *J. Ind. Inf. Integr.*, vol. 6, pp. 1–10, Jun. 2017.

[2] J. Wurm, K. Hoang, O. Arias, A.-R. Sadeghi, and Y. Jin, “Security analysis on consumer and industrial IoT devices,” in *Proc. 21st Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2016, pp. 519–524.

[3] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, “Urban computing: Concepts, methodologies, and applications,” *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 38, 2014.

[4] G. Dartmann, H. Song, and A. Schmeink, *Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things*. Amsterdam, The Netherlands: Elsevier, 2019.

[5] Y. Liang, Z. Cai, J. Yu, Q. Han, and Y. Li, “Deep learning based inference of private information using embedded sensors in smart devices,” *IEEE Netw.*, vol. 32, no. 4, pp. 8–14, Jul. 2018.

[6] H. Song and M. Brandt-Pearce, “Range of influence and impact of physical impairments in long-haul DWDM systems,” *J. Lightw. Technol.*, vol. 31, no. 6, pp. 846–854, Mar. 15, 2013.

[7] Z. Cai and Z. He, “Trading private range counting over big IoT data,” in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 144–153.

[8] S. Jeschke, C. Brecher, H. Song, and D. Rawat, *Industrial Internet of Things: Cybermanufacturing Systems*. Cham, Switzerland: Springer, 2017, pp. 1–715.

[9] L. E. Noboa, F. Lemmerich, P. Singer, and M. Strohmaier, “Discovering and characterizing mobility patterns in urban spaces: A study of manhattan taxi data,” in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 537–542.

[10] Z. Cai, X. Zheng, and J. Yu, “A differential-private framework for urban traffic flows estimation via taxi companies,” *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6492–6499, Dec. 2019.

[11] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, “City-wide traffic volume inference with loop detector data and taxi trajectories,” in *Proc. 25th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2017, pp. 1–10.

[12] Y. Zhang, Z. Guo, J. Lv, and Y. Liu, “A framework for smart production-logistics systems based on CPS and industrial IoT,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 4019–4032, Sep. 2018.

[13] V. A. Memos, K. E. Psannis, Y. Ishibashi, B.-G. Kim, and B. B. Gupta, “An efficient algorithm for media-based surveillance system (EAMSuS) in IoT smart city framework,” *Future Gener. Comput. Syst.*, vol. 83, pp. 619–628, Jun. 2018.

[14] A. P. Plageras, K. E. Psannis, C. Stergiou, H. Wang, and B. B. Gupta, “Efficient IoT-based sensor BIG data collection–processing and analysis in smart buildings,” *Future Gener. Comput. Syst.*, vol. 82, pp. 349–357, May 2018.

[15] A. Broring *et al.*, “Enabling IoT ecosystems through platform interoperability,” *IEEE Softw.*, vol. 34, no. 1, pp. 54–61, Jan. 2017.

[16] H. Song and M. Brandt-Pearce, “A 2-D discrete-time model of physical impairments in wavelength-division multiplexing systems,” *J. Lightw. Technol.*, vol. 30, no. 5, pp. 713–726, Mar. 1, 2011.

[17] H. Song, G. A. Fink, and S. Jeschke, *Security and Privacy in Cyber-Physical Systems: Foundations, Principles, and Applications*. Hoboken, NJ, USA: Wiley, 2017.

[18] C. Yin, J. Xi, R. Sun, and J. Wang, “Location privacy protection based on differential privacy strategy for big data in industrial Internet of Things,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3628–3636, Aug. 2017.

[19] A. Sarkisyan, R. Debbiny, and A. Nahapetian, “WristSnoop: Smartphone PINs prediction using smartwatch motion sensors,” in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Nov. 2015, pp. 1–6.

[20] Q. Xu, P. Ren, H. Song, and Q. Du, “Security-aware waveforms for enhancing wireless communications privacy in cyber-physical systems via multipath receptions,” *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1924–1933, Dec. 2017.

[21] Z. Cai and X. Zheng, “A private and efficient mechanism for data uploading in smart cyber-physical systems,” *IEEE Trans. Netw. Sci. Eng.*, to be published.

[22] X. Zheng, Z. Cai, and Y. Li, “Data linkage in smart Internet of Things systems: A consideration from a privacy perspective,” *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 55–61, Sep. 2018.

[23] I. Butun, P. Österberg, and H. Song, “Security of the Internet of Things: Vulnerabilities, attacks, and countermeasures,” *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 616–644, 1st Quart., 2019.

[24] X. Gong, Q.-S. Hua, L. Qian, D. Yu, and H. Jin, “Communication-efficient and privacy-preserving data aggregation without trusted authority,” in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2018, pp. 1250–1258.

[25] R. Bassily and A. Smith, “Local, private, efficient protocols for succinct histograms,” in *Proc. 47th Annu. ACM Symp. Theory Comput. (STOC)*, 2015, pp. 127–135.

[26] U. Erlingsson, V. Pihur, and A. Korolova, “RAPPOR: Randomized aggregatable privacy-preserving ordinal response,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1054–1067.

[27] S. Wang, L. Huang, P. Wang, H. Deng, H. Xu, and W. Yang, “Private weighted histogram aggregation in crowdsourcing,” in *Proc. Int. Conf. Wireless Algorithms, Syst., Appl.* Cham, Switzerland: Springer, 2016, pp. 250–261.

[28] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren, “Generating synthetic decentralized social graphs with local differential privacy,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2017, pp. 425–438.

[29] M. Bun, J. Nelson, and U. Stemmer, “Heavy hitters and the structure of local privacy,” in *Proc. 35th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, 2018, pp. 435–447.

[30] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2016, pp. 192–203.

[31] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proc. 26th USENIX Secur. Symp.*, 2017, pp. 729–745.

[32] C. Wang, C. Wang, Z. Wang, X. Ye, J. X. Yu, and B. Wang, "DeepDirect: Learning directions of social ties with edge-based network embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2277–2291, Dec. 2018.

[33] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 577–590, Sep. 2016.

[34] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, "Security and privacy preservation scheme of face identification and resolution framework using fog computing in Internet of Things," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1143–1155, Oct. 2017.

[35] L. Wang, D. Yang, X. Han, T. Wang, D. Zhang, and X. Ma, "Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, 2017, pp. 627–636.

[36] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in *Proc. IEEE 36th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2016, pp. 344–353.

[37] F. McSherry and I. Mironov, "Differentially private recommender systems: Building privacy into the netflix prize contenders," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 627–636.

[38] M. S. Hossain and G. Muhammad, "Cloud-assisted industrial Internet of Things (IIoT)–enabled framework for health monitoring," *Comput. Netw.*, vol. 101, pp. 192–202, Jun. 2016.

[39] X. Zhang *et al.*, "Incentives for mobile crowd sensing: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 54–67, 1st Quart., 2015.

[40] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Oper. Res.*, vol. 11, no. 3, pp. 399–417, Jun. 1963.

[41] (2020). *Freight Analysis Framework*. [Online]. Available: https://www.bts.gov/faf

[42] (2018). *2017 Yellow Taxi Trip Data*. [Online]. Available: https://data.cityofnewyork.us/Transportation/2017-Yellow-Taxi-Trip-Data/biws-g3hs

Xu Zheng (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Computer Science and Technology, Harbin Institute of Technology, and the second Ph.D. degree from the Department of Computer Science, Georgia State University. He is currently an Assistant Professor with the Department of Computer Science and Engineering, University of Electronic Science and Technology of China. His research areas focus on wireless networks and data security.

Zhipeng Cai (Senior Member, IEEE) received the B.S. degree from the Department of Computer Science and Engineering, Beijing Institute of Technology, and the M.S. and Ph.D. degrees from the Department of Computing Science, University of Alberta. He is currently an Associate Professor with the Department of Computer Science, Georgia State University (GSU). Prior to joining GSU, he was a Research Faculty with the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research areas focus on networking, big data, data security, and artificial intelligence. He was a recipient of the NSF CAREER Award.