This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

A Differential-Private Framework for Urban Traffic Flows Estimation via Taxi Companies

Zhipeng Cai[1], *Senior Member, IEEE,* Xu Zheng[2], *Member, IEEE,* Jiguo Yu[3,4,5,*]

***Abstract*—Due to the prominent development of public transportation systems, the taxi flows could nowadays work as a reasonable reference to the trend of urban population. Being aware of this knowledge will significantly benefit regular individuals, city planners, and the taxi companies themselves. However, to mindlessly publish such contents will severely threaten the private information of taxi companies. Both their own market ratios and the sensitive information of passengers and drivers will be revealed. Consequently, we propose in this work a novel framework for privacy-preserved traffic sharing among taxi companies, which jointly considers the privacy, profits, and fairness for participants. The framework allows companies to share scales of their taxi flows, and common knowledge will be derived from these statistics. Two algorithms are proposed for the derivation of sharing schemes in different scenarios, depending on whether the common knowledge can be accessed by third parties like individuals and governments. The differential privacy is utilized in both cases to preserve the sensitive information for taxi companies. Finally, both algorithms are validated on real-world data traces under multiple market distributions.**

I. Introduction

The emergence and dramatic development of smart transportation systems have provided unprecedentedly services for local residents. The taxi flows in urban area, jointly contributed by multiple companies, are acting as one major component in this integrated system. They can depict the fine-grained contour of urban population [1]. This type of information is critical, as the government, taxi companies, and regular residents may all rely on the knowledge to carry out customized plans [2] [3] [4]. For example, taxi companies may refer the knowledge to decide whether to deploy extra cabs in specific regions, or apply new energy automobile for long-term cost reduction [5]. The government may apply the knowledge to guide city plans and public facilities construction [6]. However, the taxi transactions could also reveal the market ratios of a company, as well as the detailed information about both drivers and passengers, which are both sensitive and may thwart the development of the smart industrial ecosystems [7]. As a result, we study in this work a novel framework for traffic flow extraction among multiple taxi companies under privacy preservation. to facilitate the discovery of taxi traffic flows in urban area via data publication,

Actually, the discovery of local taxi flows is closely related to data sharing among involved parties, *i.e.*, the shared knowledge can be aggregated and build the view for traffic flows. However, besides the privacy concerns, these taxi companies are also diverse on their market ratios and shapes, and may hold heterogeneous opinions on data sharing, ranging from building a general overview of urban traffics to gaining useful knowledge for subsequent market promotion. Then corresponding schemes for data sharing must treat all companies accordingly, and derives the knowledge and sharing scheme acceptable for all. Generally, the derivation of local traffic flows must take the benefits, the market ratios, and especially privacy concerns of involved companies into consideration.

Fortunately, the privacy issues have been investigated for multiple categories of data publication and sharing problems [8]. The Local Differential Privacy (LDP for short), as an extension of the basic differential privacy [9], allows multiple data holders to independently publish and globally merge their contents under differential privacy, However, these works all tend to apply same publication mechanism for all participants, ignoring the pervasively existed heterogeneity among participants.

As a result, we for the first time investigate a novel framework for the discovery of city-wide taxi flows via privacy-preserved data sharing among multiple companies. Our framework allows involved taxi companies to either build the traffic overview for general publics or share for their own profits. Generally, each company holds partial of taxi flows in different urban regions, and expect to gain knowledge of total traffic scales in some interested regions. To achieve the goals, companies share their scales of flows, and observe the useful knowledge within the shared results. Furthermore, our framework guarantee the differential privacy for each company, *i.e.*, companies publish perturbed traffic scales. The main challenge is to determine for each company the sharing scheme including the shared regions, and the scales of noise injected in the perturbation. In another word, this scheme should actually consider the utilities, privacy preservation, and the fairness

This work is partly supported by the National Science Foundation (NSF) under grant NOs. 1252292, 1741277, 1829674, and 1704287, the National Natural Science Foundation of China under Grant NOs. 61672321, 61771289, 61832012, 61373027, and the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 61802050).

1 Z. Cai is with Department of Computer Science, Georgia State University, 25 Park Place, Atlanta, GA, 30302, USA. Email: zcai@gsu.edu

2 X. Zheng is with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China, 611731. Email: xzheng@uestc.edu.cn, gcluo@uestc.edu.cn.

3 J. Yu is with School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong 250353, China. Email: jiguoyu@sina.com

4 J. Yu is with Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan, Shandong 250014, China.

5 J. Yu is with School of Information Science and Engineering, Qufu Normal University, Rizhao, Shandong 276826, China.

* Corresponding author: J. Yu

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

simultaneously. Specifically, the fairness is defined according to their privacy devotions, market ratios, and expectation on the observed flows. All companies should receive balanced utilities in our framework when combining the three factors.

Our framework mainly considers the sharing of local traffic flows in two scenarios. In the first case, a third party curator like the government or the dominating company leads the share among companies. One common result depicting the traffic flows in all regions will be aggregated and published to all involved taxi companies. In this case, we formulate the traffic flow sharing as a max-min problem, and derives a corresponding data sharing scheme achieving the max-min property on utilities. In the second scenario, some third parties still act as data curators. Companies will only receive partial of taxi flows in urban regions upon requests. This is the case for intra-domain knowledge exchange, where taxi companies simply exchange their information for market understanding and promotion. We formulate the data sharing problem as a coalitional game. Then a corresponding algorithm is designed to derive the scheme for data sharing.

Finally, we validate the proposed framework on real-world datasets, and multiple types of market distributions are tested. According to the experiment results, our algorithms can derive highly useful overviews for taxi flows under the strict privacy preservation. The utilities and detailed performance for both individual companies and the general publics are also evaluated and discussed in various cases. The main contribution of this paper includes:

- A formal formulation is proposed for the discovery of traffic flows among multiple heterogeneous taxi companies. Factors including requests on different regions, privacy concerns, market ratios, and general utilities are considered.
- The local differential privacy is for the first time applied to data share among heterogeneous data holders.
- A novel algorithm for data sharing to derive the general view for taxi traffic flows is designed.
- A novel algorithm for data sharing supporting knowledge exchange among taxi companies is designed.
- Both theoretical and experimental results are introduced to demonstrate the effectiveness of designed algorithms.

The remaining of this paper is organized as follow. Section 2 reviews the existing work in this area. Section 3 proposes the system settings and objectives. Section 4 and 5 introduce the algorithms designed for two scenarios, and analyze their performance respectively. Section 6 introduces the evaluation results. Section 7 concludes the paper.

II. Related Work

Building the knowledge on local traffic flows has been pivotal for urban computing and city construction [10],[11]. The subregions with heavy traffics, the generally trends of flows during weekday and weekend, and many other knowledge all contribute to various types of applications. For example, the study in [2] characterizes the urban mobility flows according to the taxi data in Manhattan, NY, and Guiyang,China. Some meaningful statistical results and observations are discussed. There are also some study focusing on other public applications, like the plan for novel bus routes [12]. Meanwhile, there are also a batch of studies investigating the applications for individuals. Some topics include the discovery of hot paths [13], and the maintaining of idle time for taxi drivers [14]. However, they mainly assume the data are hold by single party, or totally available to publics.

As for the privacy preservation toward authorized users, differential privacy has been treated as the most rigorous principle [8]. Typical studies in this field try to maximize the data utility under request privacy preservation. Multiple intermediate data structures are proposed to improve the accuracy for derived results [15]. The studies in [16] preserve sensitive data from multiple sources and allow them to be applied for model training. To extend the differential privacy to distributed scenarios, the local differential privacy (LDP for short) has been proposed [17]. The Basic RAPPOR [9] and RAPPOR [9] proposed by Google are designed to collect behaviors of users. Other knowledge including heavy hitters [18], histograms [19][20] can also be extracted under LDP. Finally, typical terminologies for privacy preservation are also considered for traffic flow publication [21] [22], such as K-anonymity, L-diversity, which provides corresponding standard for indistinguishability. However, all these studies treat all data owners equally, and ignores the complicated correlations among them. They fail to provide fairness among taxi companies in our problem.

III. System Model

In this section, we first introduce our problem formulation, including the general system inputs, the assumption on adversaries and privacy preservation, and the optimization objective of the problem.

A. System Inputs

The urban area Φ is composed of N regions $\{G_1, G_2, \cdots, G_N\}$, where each region may refer to some locally closed blocks. There are K different taxi companies $\mathscr{V} = \{V_1, V_2, \cdots, V_K\}$ serving in the area. For each V_j, it owns a traffic flow set $M_{N\times 1} = \{m_{ij}\}_{N\times 1}$, where m_{ij} indicates the ratio of taxi service in G_i provided by V_j. For example, m_{ij} could be the scale of pick-up transactions contributed by company V_j during rush hours in region G_i.

In our framework, one trusted data curator holds the scales of traffic flows for all companies. The data curators could be some dominating third parties, for example, the local government, which collects the information for the purpose of public security. The curator will determine the sharing and distributing of the traffic flows, and the supposed receivers include both taxi companies and general publics.

Taxi companies, as well as the publics, expect to gain knowledge from the traffic flows. Taxi companies apply the knowledge to support subsequent investments, aiming at promoting their markets and profits. Assume the interested regions for V_i are denoted as $Q_i = \{I_{1i}, I_{2i}, \cdots, I_{Ni}\}$, where $I_{ji} \in \{0, 1\}$ indicates whether the traffic flow in G_j is concerned by V_i. Meanwhile, the publics are interested in the General traffic flows in the whole area.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

B. Adversary Model and Privacy

In our framework, malicious data viewers are considered as the adversaries, which could be either business rivals or malicious individuals. These viewers are honest-but-curious. They will collect the flow information published by the data curator, and extract necessary knowledge on the local traffics. Meanwhile, these data viewers also tries to infer the private information beneath the flows. Therefore, we apply the Differential Privacy (DP for short) to measure the preservation on both traffic flow scales and every single taxi transaction for each company. DP is a typical measurement for privacy preservation of aggregated information among multiple data owners. The formal definition of DP is given in definition 1.

Definition 1 (Differential Privacy [23]): Assume the set of taxi transactions from taxi company V_i to be D_i, and D_i' is another service set that differs with D_i on just one transaction. Then an algorithm or a data sharing mechanism A for traffic flow discovery satisfies ϵ-differential privacy (ϵ-DP) where $\epsilon \geq 0$, if and only if

$$\forall y \in Range(A) : Pr[A(D_i) = y] \leq e^{\epsilon} Pr[A(D_i') = y],$$

where Range(A) denotes the set of all possible outputs of A.

Generally, DP assures the adversaries will not learn significant information on single transaction from the observed results, even if they already own strong prior knowledge on the remaining part of the taxi service set. A larger ϵ indicates the companies are more tolerable and less sensitive. The privacy budget of V_i is denoted as ϵ_{max}. For simplicity, we will use terms privacy preference and privacy budgets interchangeably in our study. The following Laplace mechanism guarantees ϵ-differential privacy [23].

Theorem 1: For any function $A(\cdot) : G \rightarrow R$, the randomized algorithm F provides ϵ-differential privacy when

$$F = A(D_i) + Lap(\frac{\Delta A}{\epsilon}), \tag{1}$$

where $Lap(\frac{\Delta A}{\epsilon})$ follows Laplace distribution with scaling factor $\frac{\Delta A}{\epsilon}$, and ΔA refers to the global sensitivity of function A: $\max |A(D_i) - A(D_i')|, \forall D_i$

Finally, the composition principles for differential privacy are given:

Theorem 2 (Parallel Composition [24]): Let $\{d_{i1}, d_{i2}, \cdots, d_{iK_i}\}$ be disjoint transaction sets held by taxi company V_i, and A_is be a set of K_i data share mechanisms each providing ϵ_i-differential privacy. Then applying all A_is to their corresponding sets d_{ij}s can guarantee a $\max\{\epsilon_i\}$-differential privacy for V_i.

C. Utility for Taxi Companies and Publics

Taxi companies participate in the data sharing to derive valuable knowledge for their market promotion. Therefore, they expect to extract knowledge highly consistent with the ground truth, which leads to unbiased decisions. Assume $M_i' = \{M_{1i}', M_{2i}', \cdots, M_{Ni}'\}$ to be the derived traffic flows for V_i. Then the relative accuracy for V_i is $\sum_{l=1}^{N}(|\sum_{j=1}^{K} m_{lj} - M_{li}| \cdot I_{li})$, which is the summation of difference between the observed values and ground truth in interested area. Considering that our framework is designed to preserve the private transactions, each taxi company will likely receive perturbed scales. Furthermore, the expectation of observed traffic scales is identical with the ground truth, as the expected value of Laplace mechanism is zero. Therefore, we apply the following term to indicate the utility for each company:

$$P_i = \sum_{j=1}^{N}\sum_{k=1}^{K}(\frac{m_{jk}\epsilon_j^2}{\sum_{k=1}^{K} m_{jk}}), \tag{2}$$

where $\frac{1}{\epsilon_j^2}$ indicates the variance on the scale of traffic flows shared by V_j. Generally, a larger ϵ_j means the taxi company is less sensitive on its transactions, and therefore generating results with with less variance $\delta \propto \frac{1}{\epsilon_j^2}$. Then the overall utility of V_i is the weighted summation of ϵ_j^2 in all interested regions, indicating the total variance in observed results.

IV. Solution for Common Urban Traffic Sharing

A. Background and Preliminaries

In the first scenario, a leading third party acts as the trusted data curator V_0. V_0 has access to the accurate scales of all traffic flows, and collects the requests and privacy preference from all companies. The privacy preference refers to the maximum ϵ_{max} allowed by each company. V_0 selects the proper privacy settings applied for taxi companies, and perturbs the scales with Laplace mechanism according to the settings. Finally, the obfuscated flows in each regions are accumulated and published. This is the case where both taxi companies and publics are allowed to learn meaningful knowledge from the data share.

As companies own heterogeneous markets and interests on traffic flows, V_0 is expected derive a sharing scheme acceptable for all companies. More specific, the scheme should guarantee the fairness among companies, and jointly consider the utility and the devotion:

$$\max \min_i P_i = \sum_{j=1}^{N}\sum_{k=1}^{K}(\frac{m_{jk}\epsilon_j^2}{\sum_{k=1}^{K} m_{jk}}) \cdot \frac{1}{\sum_{j=1}^{N} m_{ji} \cdot \epsilon_i^2}, \tag{3}$$

which tries to maximize the minimum relative benefits P_is. This principle provides a rational benefits for both individual companies and publics. According to equation 3, the benefit for each company is influenced by privacy settings provided by correlated companies in interested regions. More specific, we can convert the benefits according to the correlations among companies, by introducing $\alpha_{ik} = \sum_{j=1}^{N}(\frac{I_{ji}m_{jk}}{\sum_{l=1}^{K} m_{jl}})$, and $\beta_i = \sum_{j=1}^{N} m_{ji}$.

$$P_i = (\alpha_{i1}\epsilon_1^2 + \alpha_{i2}\epsilon_2^2 + \cdots + \alpha_{iK}\epsilon_K^2) \cdot \frac{i}{\beta_i \epsilon_i^2} \tag{4}$$

According to equation set 4, we observe that $\alpha_{ik} > 0$ if and only if there is at least one region G_j concerned by V_i, and V_k shares market in G_j.

We also introduce a graph structure called ***Benefiting Graph*** $G_B = \{V, E\}$, where V refers to all companies,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

and $E = \{e_{ij}\}, i, j \in \{1, 2, \cdots, K\}$. We assume e_{ij} exists when $\alpha_{ji} > 0$,and G_B is a directed graph. e_{ij} means the traffic flows shared by V_i contributes to the utility of V_j, and larger privacy setting will improve the utility. Furthermore, a strongly connected component of G_B indicates a set of vertex $C_L = \{V_{i_1}, V_{i_2}, \cdots, V_{i_L}\}$, such that there is a path connecting any V_{i_j}, V_{i_k} inside. We also propose an improved definition for fairness among companies according to the graph.

Definition 2: For an arbitrary set of companies $\{V_{i_1}, V_{i_2}, \cdots, V_{i_L}\}$ in a strongly connected component, the privacy settings $\{\epsilon_{i_1}, \epsilon_{i_2}, \cdots, \epsilon_{i_L}\}$ achieves a balanced state when no other assignments $\{\epsilon'_{i_1}, \epsilon'_{i_2}, \cdots, \epsilon'_{i_L}\}$ can provide $P_i < P'_i \leq P'_j < P_j$, $P_k = P'_k, \forall k \neq i, j$ for at least one pair of i, j, or increase the utility for all companies under the given privacy budget.

Definition 2 generally ensures companies receive fair benefits according to their devotion, while the budgets are sufficiently applied to maintain good overall utility for both companies and other third parties. This new definition of fairness can further improve the balance among companies by overcoming the unnecessary conservation on privacy budgets.

B. Common Knowledge Acquisition Algorithm

This part introduces in details the algorithm designed for Common Urban Traffic Knowledge Sharing. The algorithm takes the market ratios, the interested regions, and the privacy budgets of all companies as inputs, determines the privacy settings ϵ_is applied by these companies in the data sharing, and outputs the privacy-preserved scales of local traffics for all regions. For clarity, the algorithm is named as ***Common Knowledge Acquisition Algorithm*** (CKAA for short).

CKAA is mainly composed of four parts. It first derives the benefit graph and the benefiting equation set according to the inputs. Then CKAA searches for all strongly connected components in the graph. In the third phase, CKAA determines the privacy settings for all companies according to the benefiting formulas and the strongly connected components. The privacy settings achieve the requirements in Definition 2. Finally, CKAA perturbs the scale of traffic flows in each region for each company and publishes the final results. Next we provide the details for each phase.

In the first phase, CKAA initializes the benefit graph G_B from the inputs. Every company is represented by a vertex in G_B. For simplicity, we utilize V_i to denote both the company and its corresponding vertex in G_B. Furthermore, there is a directed edge from V_i to V_j,$\forall i, j$, if and only if V_i owns markets in any of regions requested by V_j. We also construct the benefiting formulas like equations 4, which is also derived from the inputs by $\alpha_{ik} = \sum_{j=1}^{N}(\frac{I_{ji} m_{jk}}{\sum_{l=1}^{K} m_{jl}})$, and $\beta_i = \sum_{j=1}^{N} m_{ji}$.

In the second phase, CKAA derives all strongly connected components (SCC for short) from G_B, denoted as $C_1, C_2, \cdots, C_P$. We apply the Kosaraju's algorithm [25] for the detection of $C_1, C_2, \cdots, C_P$. The main idea of the algorithm is to first reverse all edges in G_B, and applies a post-order Depth First Search (DFS for short) to traverse all vertices. Then the algorithm pops up the top vertex from the stack, and applies a second DFS starting from the top vertex in the original G_B. After the DFS, a strong connected component will be derived, and corresponding vertices are labeled. The algorithm iteratively pops up the unlabeled top vertices from the stack, and derives subsequent SCCs. The whole algorithm terminates when all vertices are labeled. After this phase, CKAA partitions all companies into multiple groups as $C_1, C_2, \cdots, C_P$.

In the third phase, CKAA determines the privacy settings for companies. To achieve this, CKAA first checks the inter-edges among all components $C_1, C_2, \cdots, C_P$, and extracts all components with only out-edges, marked as $C_{o1}, C_{o2}, \cdots$. Then for each C_{oi}, CKAA sets the benefits of all companies inside to be identical. In this case, there are $|C_{oi}|$ privacy settings to be determined, and $|C_{oi}| - 1$ equations, where $|\cdot|$ denotes the cardinality of the set. CKAA iteratively sets the privacy setting of each company to be maximum, and derives the settings for remaining companies by solving these equation sets. It stops when the all derived results are no larger than the maximum budget. Subsequently, CKAA updates the determined settings in G_B, and deletes all inter-edges whose source vertices have their privacy settings determined. After the updating, CKAA starts over and searches for new components with only out-edges. The third phase ends when the settings for all companies are decided.

In the fourth phase, CKAA perturbs the traffic flows with assigned settings to generate the final scales of local traffic for all regions. For each company V_j and region G_i,

$$m'_{ij} = m_{ij} + Lap(\epsilon_j), \tag{5}$$

where m'_{ij} is the published scale of traffic flows for V_j in region G_i. Finally, CKAA publishes a perturbed city-wide view for the scales of traffic flows, where the size of flows in an arbitrary region G_i is

$$M_i = \sum_{j=1}^{K} m'_{ij}. \tag{6}$$

C. Performance Analysis

In this part, we first prove that our algorithm can achieve the requested fairness in Definition 2. Secondly, CKAA is proved to guarantee the ϵ-differential privacy for all companies.

Fairness:

As is shown in Definition 2, the fairness expects both the fully utilization of privacy budgets and the balanced benefits among companies. We can easily achieve the first requirement as CKAA always tries to make full use of the privacy budget, and any other assignment of privacy settings will lead to the violation of privacy for at least one company. For the second conclusion, we can prove it with the following theorem.

Lemma 1: Assume $\{V_{i_1}, V_{i_2}, \cdots, V_{i_L}\}$ to be a strong connected component in G_B. Then the max-min principle $\max\min_{k \leq L} P_{i_k}$ is achieved when $P_{i_1} = P_{i_2} = \cdots = P_{i_L}$.

Proof: We prove the theorem by contradiction.

Denote $\{\epsilon_{i_1}, \epsilon_{i_2}, \cdots, \epsilon_{i_L}\}$ to be the privacy assignment for companies in one strongly connected component, and there

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

is another assignment $\{\epsilon'_{i_1}, \epsilon'_{i_2}, \cdots, \epsilon'_{i_L}\}$ making some $P'_{i_j} \neq P_{i_k}$ and $\max\min P'_{i_k} > \max\min P_{i_k}$. We say two companies share markets when their traffic flows in some regions are requested by each other. Then one of the following facts holds for $\{\epsilon'_{i_1}, \epsilon'_{i_2}, \cdots, \epsilon'_{i_L}\}$:

1) There are multiple companies with minimum benefit, and each of them, saying V_{i_j}, shares markets with at least one company with larger benefit, denoting by V_{i_k}. If $\epsilon_{i_k} \leq \epsilon_{max}$, we can simply increase ϵ_{i_k} and achieve new benefits:

$$P'_{i_k} = P'_{i_l}, \tag{7}$$

where both benefits are larger than P'_{i_j}.

2) There are multiple companies with minimum benefit, and only partial of them share markets with companies with higher benefits. Then we can first apply the strategy in case 1 to increase the benefits for these companies. All remaining companies slightly decrease their budgets to make their benefits identical with other companies affected in the first step. Then we can still achieve a new assignment for privacy settings that increases the benefits for all companies.

3) There are multiple companies with minimum benefit, and none of them shares markets with companies with higher benefits. This is contradicted with the facts that these companies fall in the same strongly connected component.

Therefore, when the benefits for companies in one strongly connected components are non-identical, we can always derive another assignment to increase the minimum benefit. The proof is completed. ■

We can extend the proof to all components, and CKAA follows the requested fairness in definition 2. Furthermore, as CKAA tries to make full use of privacy budgets under the fairness principle, it also meet the maximization of utility in definition 2. Generally, we have:

Theorem 3: The privacy settings derived by CKAA can maximize the minimum benefits for companies in all SCCs, and the utility is also maximized under the principle of fairness.

Privacy Preservation:

Now we analyze the performance of CKAA on privacy preservation. The main steps for privacy preservation are located in the fourth phase, where each company perturbs its scale of traffic flows in each region according to the assigned privacy setting. The perturbed results are accumulated and construct the final outputs. The global sensitivity of the function is 1 as the existence or absent of a transaction will affect the scale of traffic flows by 1. The following theorem demonstrate the privacy preservation of CKAA.

Theorem 4: CKAA guarantees ϵ_{max}-differential privacy on the traffic flows for all involved taxi companies.

CKAA achieves ϵ_{max}-differential privacy in each region, as Laplace mechanism is adopt and the max budget is bounded for each company. Furthermore, as the traffic scales are disjoint in different regions, CKAA can guarantee the ϵ_{max}-DP in all regions according to the parallel principle in theorem 2.

V. Solution for Local Traffic Exchange

A. Background and Preliminaries

In the second scenario, there is also a trusted data curator V_0 maintaining the markets, interests, and privacy preference for taxi companies. Unlike the first scenario, V_0 implements a local exchange of knowledge among taxi companies. This data curator will assign the privacy setting for each company in each region, and determines a set of company-region pairs indicating the companies and corresponding regions that a taxi company should exchange its markets with. This is the case where companies locally derive knowledge on traffic flows via data sharing, and no common knowledge will be published. As companies concentrate on their own profits, the main objective is to keep the sharing rational and stable among them. Furthermore, the flows in each region could be perturbed once an shared multiple times, which further facilitate the exchange among companies.

We formulate the interactions among companies as a coalitional game, where all companies form the player set $\mathscr{V}$. The strategies $\mathscr{X}_V$ are composed of all feasible sharing among companies, which is the combination of shared region and the privacy setting applied between two arbitrary companies. The merit of a sharing is correlated with the shared markets and the accuracy, which is denoted as $\frac{m_{ij}}{\sum_{k\leq K} m_{kj}} \cdot \epsilon_i^2$.

Based on the formulation of the coalitional game, the data sharing actually indicates a subset of exchanges $\mathscr{X}_V^E$ from $\mathscr{X}_V$ for companies. The actual utility P_j^0 for company V_j is correlated with the summation of all retrieved traffic flows during the coalition. Then the accuracy of the results is applied to evaluate the utility as companies expect unbiased and reliable knowledge. We apply the typical core solution to evaluate the performance of a coalitional in our framework.

Definition 3: The ***Core*** for a coalitional game Ω is a set $\mathscr{X}_V^E \subset \mathscr{X}_V$, such that there are no other subsets $\mathscr{X}_V^{E'}$s providing larger benefits P_is for all involved companies.

Before introducing details of the proposed algorithm, we briefly introduce the ***Dependent Graph*** applied by the process. The ***Dependent Graph*** $G_D = \{R, E\}$ is a directed graph, where R includes all companies, *i.e.*, $|R| = K$. Assume r_i to stand for company V_i. Then there is a directed edge e_{li} between vertices r_i and an arbitrary r_l when company V_l requests the traffic market in some regions owned by V_i. Furthermore, we set multiple weights on e_{li} as $\{\frac{m_{1i}}{\sum_{p\leq K} m_{1p}} \cdot \epsilon_{max}^2, \frac{m_{2i}}{\sum_{p\leq K} m_{2p}} \cdot \epsilon_{max}^2, \cdots, \frac{m_{Ni}}{\sum_{p\leq K} m_{Np}} \cdot \epsilon_{max}^2\}$, indicating the maximum benefit V_i can provide by safely sharing its market in each region. The dependent graph is diverse from benefiting graph on the definition of weights to facilitate a fine-grained coalition.

B. Local Traffic Exchange Algorithm

In this part, we introduce the process of each step for EKAA. The main idea of EKAA is to iteratively derive currently best interactions for sharing, and updates the dependent graph G_D accordingly.

Initially, EKAA sorts weights on all edges in descending order. Edges with identical weights are stored in one slot in the list. Then starting from the edge e_{ij}s in the first slot, EKAA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

sets $w_{ij} = w_{i'j'}$, where $e_{i'j'}$ is the edge succeeding to e_{ij} in the sorted list. Edges in last slot keeps their weight stable.

In each iteration, EKAA derives a top-order circle from G_D. EKAA first deletes the vertices with only in-edges in G_D. The weight of each edge is set to be the maximum weight among its candidate set. EKAA starts from a random vertex. EKAA searches for the edge with maximum weights, saying $e_{i_1 i_2}$. Then EKAA traverses to next vertices r_{i_3}, where $w_{i_2 i_3}$ is the maximum weights among all out edges of r_{i_2}. EKAA repeats the procedure until the most recent vertex r_{i_k} refers to a company once appeared in the path, *i.e.*, partial of visited vertices form a circle Cr_l. Vertices in the circle, together with the corresponding regions contributing the weights on each edge, are included in the final outputs. Meanwhile, all the weights in selected edges are removed from the candidate sets, and the second largest weights in the sets are applied. The whole edges are removed when all non-zero weights are included in the outputs. Then EKAA starts over with removing vertices with only in-edges in G_D, and selects another vertex to search for next top-order circle.

The searching process terminates when all edges have been deleted or all vertices have been removed. EKAA outputs a set of circles $\{Cr_1, Cr_2, \cdots\}$, where each circle contains a list of companies, and the regions and privacy settings applied for sharing among them. Then EKAA perturbs the scale of traffic flows for each company in each region according to their weights. The noise is also introduced by the Laplace mechanism:

$$m'_{ij} = m_{ij} + Lap(\epsilon_{ij}), \tag{8}$$

where m'_{ij} is the published market size for V_j in region G_i, and ϵ_{ij} is the privacy setting for V_j in region G_i. Finally, EKAA delivers the obfuscated markets to corresponding companies in circles.

C. Analysis

In this section, we first prove the stability of the proposed algorithm. EKAA can derive a core set for coalition among participated companies.

The stability of a coalitional game indicates the achieved scheme for data exchange will be convincing and accepted by all taxi companies. In the typical coalitional game, the core set is usually adopt to evaluate the stability of the scheme, which is given in definition 3. We prove that the proposed algorithm can generate a core set for data sharing among participants.

Theorem 5: The cooperation scheme $\{Cr_1, Cr_2, \cdots\}$ derived by EKAA is a core set for all involved participants.

Proof: We prove the theorem by contradiction.

Assume there is another sharing scheme $\{Cr'_1, Cr'_2, \cdots\}$ providing better benefits for all involved participants. Then all sharing schemes must forms a circle. Otherwise, the participant at the tail of the path may abandon the cooperation while its benefits remains the same. Meanwhile, one of the following conditions must hold for $\{Cr'_1, Cr'_2, \cdots\}$:

1) The pairs of sharing among companies are identical with $\{Cr_1, Cr_2, \cdots\}$, but participants receives higher utilities as the selected regions are different in the scheme. This is contradicted with the fact that EKAA will priorly select the region with maximum utility.

2) There are some circles unique in $\{Cr'_1, Cr'_2, \cdots\}$, where none of its sharing pairs appeared in the original scheme. However, this is also impossible since EKAA will iteratively derives feasible circles. In this case, the unique circles will also be constructed in EKAA, as their participant-region-participant pairs are all unique.

3) There are some circles in $\{Cr'_1, Cr'_2, \cdots\}$ partially overlapping with circles in the original scheme. Assume Cr'_i to be the circle, and it is different from one circle in original scheme starting from vertex r'_{i_k}. Furthermore, the original next-hop company for r'_{i_k} is not included in $\{Cr'_1, Cr'_2, \cdots\}$. Then the utilities for r'_{i_k} will be decreased as its next-hop company is different from the one in the original scheme, which provides maximum utility.

Generally, EKAA can achieve the same utility for involved participants whenever there is a better scheme, which means no schemes can increase the utilities for all of them. Therefore, EKAA can derive a core set for the coalition game. ■

VI. Experiments

This section evaluates the effectiveness of our algorithms through extensive experiments. It first introduces the applied dataset and basic settings. Then multiple aspects of evaluation results are presented, including the overview of published traffic flows, and the effectiveness of proposed algorithms.

A. Datasets and Settings

Datasets. We apply the dataset recording the taxi transactions in New York during the year 2017 [26]. The dataset includes approximately 1.1 billion of taxi transactions, each recording the pick-up and drop-off locations of the tour. Specifically, we focus on those flows starting from or ending in Manhattan, and the goal of the data sharing is to publish the total number of transactions in each subregion in the district.

Parameters. In our evaluation, the whole area is partitioned into 15 regions, each combining a group of nearby blocks. We conduct the experiment in different settings by varying mainly three parameters, including the privacy budgets ϵ_{max}, the request ratio δ, and the size of companies K. The request ratio refers to the proportion of regions requested by compnanies. The default value for privacy budget is 10, the request ratio is set to 0.3, and the company size is 10. We assume companies share the markets in the whole area. Four types of market sharing modes are considered. These modes refer to some of the real market sharing in local business, depending on whether market-dominating companies exist and how remaining companies share the market. In ***Dominating Mode***, the dominating companies own majority of markets in the whole area, and the following companies share the remaining parts. In ***Stage Mode***, the large company still owns a relative large portion of markets in the area. The medium companies constitute moderate sizes of traffic flows, and small companies own the remaining minor parts. In ***Overlapping Mode*** and ***Non-overlapping Mode*** (Non-OLP for short), companies own similar markets in the whole

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics

area, They are mainly differ in whether many taxi companies contribute to the flows in each region.

Algorithms To the best of our knowledge, no previous works have studied the data sharing among multiple companies considering the fairness, market ratios, requests, and privacy simultaneously. Therefore, we implement and evaluate some baseline algorithms: ***Global Max-Min*** method: the utilities of all companies are set to be identical. ***Market-Based*** method: the fairness is evaluated based on the market ratio. ***Random Response*** (R-Response for short) method: follow the same strategy for scheme generation in EKAA, while applying randomly determined regions instead of the assigned one. ***Company Level*** method: the market of each company will be selected only once under EKAA.

Metrics Three metrics are applied to evaluate the performance of proposed algorithms, including the utilities, the average privacy settings, and the ratio of responded requests. The general utilities and the individual utilities estimate the accumulated utilities in the requested regions for companies, *i.e.*, the weighted summation of privacy settings in regions requested by at least one company. The average privacy settings indicate the devotion of companies in the first scenario, and the ratio of responded requests indicates the portion of requests for each company, which are selected and responded in the second scenario.

B. Performance on Observed Traffic Flows

This part evaluates the performance on the observed traffic flows in both scenarios. We will show the overall scales for traffic flows in each region. The settings in this part follow the default ones. The observed results are given in Fig.2, and fig.1 introduces the partition of local area in Manhattan, NY.

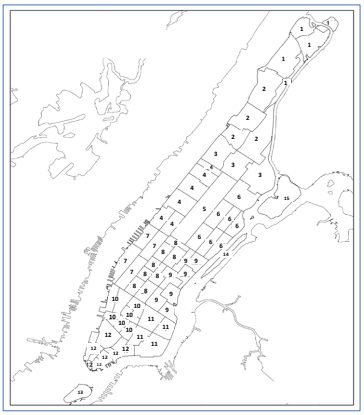
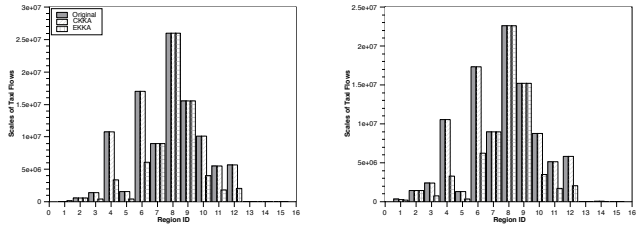


Fig. 1: Region Partition

According to the results, both algorithms provide meaningful results. Especially for the CKAA, results are highly consistent with the ground truth on contours and scales.

We also compare the difference between the observed results and the ground truth in Table. I and II. The difference is calculated as the summation of relative difference in all regions. As we see, our algorithms only constitute small scales of difference, which is less than 1 in most cases.



(a) Pick-Up (b) Drop-Off

Fig. 2: Observed Taxi Flows in Dominating Market Mode

TABLE I: Average Variance on Pick-up Traffic Flows under Various Market Modes

Market Mode	Dominating	Stage	Overlapping	Non-OLP
CKAA	0.401	0.266	0.518	0.2835
EKAA	8.457	1.074	2.120	0.355

TABLE II: Average Variance on Drop-off Traffic Flows under Various Market Modes

Market Mode	Dominating	Stage	Overlapping	Non-OLP
CKAA	0.347	0.141	0.006	0.142
EKAA	5.684	1.125	1.994	0.429

C. Basic Performance

In this part, we further investigate the performance of both algorithms under various parameter settings. The objective is to validate how the changing of parameters will influence the results for data sharing.

We evaluate the impact of privacy budgets. This metric indicates the degree to which companies allow to conceal their private information. Generally, larger privacy budgets indicate that companies are more flexible on their sensitive transactions, and will provide more accurate results. We varies the privacy budgets as 5, 10, 15, 20, 25. The results are given in Fig. 3 and Table. III.

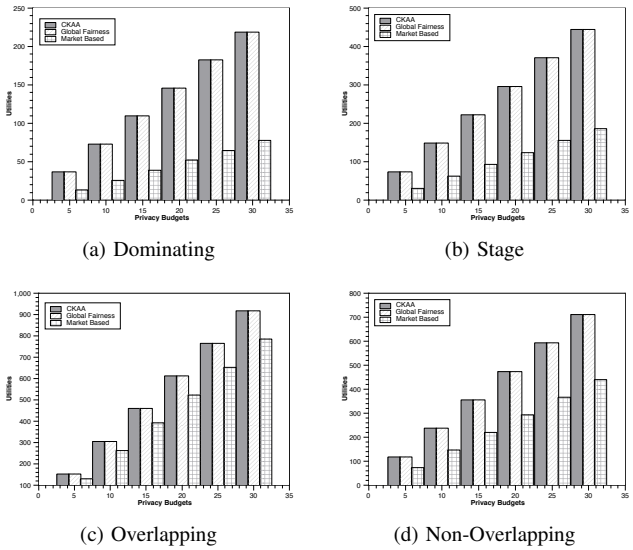
TABLE III: Ratio of Responded Requests in Scenario 2

Market Mode	Dominating	Stage	Overlapping	Non-OLP
EKAA	0.831	0.853	0.942	0.836
R-Response	0.831	0.853	0.942	0.8363
Company Level	0.327	0.171	0.259	0.263

As we see, all methods reveal improvements on the total utilities, due to more privacy budgets devoted into the sharing. We also find that CKKA algorithm significantly outperforms the market-based method. Meanwhile, our method is also more reasonable as it takes both the markets and the requests into consideration. CKAA also achieves similar performance with the global max-min method, which indicates that all companies are included in one single strongly connected component.

The performance of EKKA is also significantly better than the company-level cooperation method. which means the duplicated exchange of noisy data can significantly facilitate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2911697, IEEE Transactions on Industrial Informatics



(a) Dominating (b) Stage

(c) Overlapping (d) Non-Overlapping

Fig. 3: Utilities under Various Privacy Budgets in Scenario 1

the cooperation among companies, while keeping the privacy preserved. Both the utilities and the responded ratios are improved due to the pervasive coalition. Furthermore, EKKA also slightly outperforms the randomized response method. The reason is that EKKA will always respond with best available markets to facilitate the utilities.

We also consider the heterogeneous request ratios for companies, which means companies own different opinions on how to promote their markets. The overlapping mode is applied to mitigate the influence of different market ratios. Companies are partitioned into lazy (Company 1-3), moderate (Company 4-6), and active (Company 7-10). each with 0.3, 0.5, 0.7 request ratios. According to the results, our algorithm can guarantee the fairness among companies (*i.e.*, 5.111) while the market-based method fails to achieve a rational result (*i.e.*, 4.957, 1.832, 10.360, 4.491, 6.089, 3.782, 4.467, 7.304, 5.401, 9.789).

VII. Conclusion

In this paper, we investigate the problem of privacy-preserved traffic flow sharing among multiple taxi companies. The outputs will bring valuable knowledge for subsequent construction in urban district. Two scenarios are investigated based on whether the final outputs are formed as a general view, and corresponding algorithms are proposed respectively, where the first algorithm guarantees the fairness among companies, and the second algorithm achieves a stable scheme for cooperation. Both algorithms provide differential privacy for involved companies. Moreover, extensive evaluation results demonstrate the performance of our solutions. Potential future research includes the cases where the requests are also sensitive.

Acknowledgement

This work is partly supported by the National Science Foundation (NSF) under grant NOs. 1252292, 1741277, 1829674, and 1704287, the National Natural Science Foundation of China under Grant NOs. 61672321,61771289, 61832012, 61373027, and the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 61802050),

References

[1] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[2] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 1, ACM, 2017.

[3] M. G. Speranza, "Trends in transportation and logistics," *European Journal of Operational Research*, vol. 264, no. 3, pp. 830–836, 2018.

[4] H. Song, R. Srinivasan, T. Sookoor, and S. Jeschke, *Smart cities: foundations, principles, and applications*. John Wiley & Sons, 2017.

[5] Y. Zou, S. Wei, F. Sun, X. Hu, and Y. Shiao, "Large-scale deployment of electric taxis in beijing: A real-world analysis," *Energy*, vol. 100, pp. 25–39, 2016.

[6] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," *Journal of Transport Geography*, vol. 43, pp. 78–90, 2015.

[7] Y. Sun and H. Song, *Secure and Trustworthy Transportation Cyber-Physical Systems*. Springer, 2017.

[8] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[9] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, ACM, 2014.

[10] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.

[11] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart iot systems: A consideration from privacy perspective," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 55–61, 2018.

[12] S. P. Chuah, H. Wu, Y. Lu, L. Yu, and S. Bressan, "Bus routes design and optimization via taxi data analytics," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 2417–2420, ACM, 2016.

[13] L. Zheng, D. Xia, X. Zhao, L. Tan, H. Li, L. Chen, and W. Liu, "Spatial–temporal travel pattern mining using massive taxi trajectory data," *Physica A: Statistical Mechanics and its Applications*, vol. 501, pp. 24–41, 2018.

[14] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1202–1212, 2017.

[15] C. Lin, P. Wang, H. Song, Y. Zhou, Q. Liu, and G. Wu, "A differential privacy protection scheme for sensitive big data in body sensor networks," *Annals of Telecommunications*, vol. 71, no. 9-10, pp. 465–475, 2016.

[16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, ACM, 2016.

[17] C. Lin, Z. Song, H. Song, Y. Zhou, Y. Wang, and G. Wu, "Differential privacy preserving in big data analytics for connected health," *Journal of medical systems*, vol. 40, no. 4, p. 97, 2016.

[18] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," in *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 435–447, ACM, 2018.

[19] S. Wang, L. Huang, P. Wang, H. Deng, H. Xu, and W. Yang, "Private weighted histogram aggregation in crowdsourcing," in *International Conference on Wireless Algorithms, Systems, and Applications*, pp. 250–261, Springer, 2016.

[20] Q. Miao, W. Jing, and H. Song, "Differential privacy–based location privacy enhancing in edge computing," *Concurrency and Computation: Practice and Experience*, p. e4735, 2018.

[21] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing nyc taxi data: Does it matter?," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 140–148, IEEE, 2016.

[22] Z. Zhang, A. Tong, L. Zhu, M. Chen, and P. Su, "An anonymous scheme for current taxi applications," in *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing*, pp. 168–172, IEEE, 2016.

[23] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, pp. 265–284, Springer, 2006.

[24] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 627–636, ACM, 2009.

[25] M. Sharir, "A strong-connectivity algorithm and its applications in data flow analysis," *Computers & Mathematics with Applications*, vol. 7, no. 1, pp. 67–72, 1981.

[26] "2017 yellow taxi trip data," in *https://data.cityofnewyork.us/Transportation/2017-Yellow-Taxi-Trip-Data/biws-g3hs*, 2018.

Zhipeng Cai received his PhD and M.S. degree in Department of Computing Science at University of Alberta, and B.S. degree from Department of Computer Science and Engineering at Beijing Institute of Technology. Dr. Cai is currently an Associate Professor in the Department of Computer Science at Georgia State University. Prior to joining GSU, Dr. Cai was a research faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology. Dr. Cai's research areas focus on networking, big data, data security, and artificial intelligence. Dr. Cai is the recipient of an NSF CAREER Award.

Xu Zheng received his PhD, M.S. and B.S. degree from School of Computer Science and Technology at Harbin Institute of Technology. Mr. Zheng received his second PhD degree from Department of Computer Science at Georgia State University. Mr. Zheng is currently an assistant professor in the Department of Computer Science and Engineering at University of Electronic Science and Technology of China. Mr. Zheng's research areas focus on wireless network and data security.

Jiguo Yu received his Ph.D. degree in School of mathematics from Shandong University in 2004. He became a full professor in the School of Computer Science, Qufu Normal University, Shandong, China in 2007. Currently he is a full professor in Qilu University of Technology (Shandong Academy of Sciences), Shandong Computer Science Center (National Supercomputer Center in Jinan), and a professor in School of Information Science and Engineering, Qufu Normal University. His main research interests include privacy-aware computing, wireless networking, distributed algorithms, peer-to-peer computing, and graph theory. Particularly, he is interested in designing and analyzing algorithms for many computationally hard problems in networks. He is a senior member of IEEE, a member of ACM and a senior member of the CCF (China Computer Federation).