Towards Privacy-Preserving Data Trading for Web Browsing History

Hui Cai
Shanghai Jiao Tong University
Shanghai, China
carolinecai@sjtu.edu.cn

Fan Ye
Stony Brook University
Stony Brook, New York, USA
fan.ye@stonybrook.edu

Yuanyuan Yang
Stony Brook University
Stony Brook, New York, USA
yuanyuan.yang@stonybrook.edu

Yanmin Zhu*
Shanghai Jiao Tong University
Shanghai, China
yzhu@sjtu.edu.cn

Jie Li
Shanghai Jiao Tong University
Shanghai, China
lijie@cs.tsukuba.ac.jp

ABSTRACT

The trading of social media data has attracted wide research interests over years. Especially the trading for web browsing histories probably produces tremendous economic value for data consumers when being applied to targeted advertising. However, the disclosure of entire browsing histories, even in form of anonymous datasets poses a huge threat to user privacy. Although some existing solutions have investigated privacy-preserving outsourcing of social media data, unfortunately, they neglected the impact on the data consumer's utility. In this paper, we propose *PEATSE*, a new Privacy-prEserving dAta Trading framework for web browSing historiEs. It takes users' diverse privacy preferences and the utility of their web browsing histories into consideration. *PEATSE* perturbs users' detailed browsing times on released browsing records to protect user privacy, while balancing the privacy-utility tradeoff. Through real-data based experiments, our analysis and evaluation results demonstrate *PEATSE* indeed achieves user privacy protection, the data consumer's accuracy requirement, and truthfulness, individual rationality as well as budget balance.

CCS CONCEPTS

• **Security and privacy → Economics of security and privacy**.

KEYWORDS

web browsing history, privacy-preserving, data trading.

1 INTRODUCTION

The web browsing history refers to the set of web pages a user ever visited in previous online activities, and usually includes titles of web pages and corresponding URLs [5]. The emergence of big data era generates petabytes of data [1] per day like the web browsing history, which includes social media data (such as tweets), financial, health and video data. Furthermore, these data have tremendous usage for data consumers to study users' preferences, and further deliver targeted advertising based on the inferred preference. For example, Nestle corporation achieved a 52% lift in engagement rate in 2015 compared to the overall performance in 2014 when Behavior targeting on Twitter [4] was leveraged to infer the user audience preference.

In addition to the web browser and third-party trackers, Internet Service Providers (ISPs) such as Verizon and AT&T also have full access to individuals' web browsing histories. However, some countries in European Union have regulated a new data protection legislation, and online trackers or ISPs will face serious punishment when violating users' privacy according to General Data Protection Regulation (GDPR). Moreover, when data contributors gradually recognize the economic value of browsing histories and potential consequences of privacy disclosure [22], they are likely to allow only a trusted data broker to access their private data provided that they can receive reasonable money compensation in turn [10]. Consequently, the largest data broker of the nine typical data markets by FTC's survey [3], Acxiom [4], is allowed to access personal users' browsing histories, and then sells the related user dataset to large corporations like Microsoft or Oracle. Thus, data privacy is monetized as a commodity to be bought and sold in practical data markets [14][15].

However, the disclosure of entire web browsing histories even after anonymization still poses a serious threat to user privacy [27]. For example, the attacker can match a victim's public social media status (*i.e.*, Facebook moments or Twitter tweets) with the given user record of the received user dataset, and further acquire this victim's other sensitive information such as health status. Consequently, privacy leak [25] may result in serious consequences (*e.g.*, phishing, identify theft). In addition, the trading of perturbed private datasets has more appealing benefits than single queries, where the traded private dataset facilitates data consumers' analysis, and mitigates the additional work of dividing the original task into multiple queries. Therefore, it is crucial to design an efficient privacy-preserving trading mechanism for entire web browsing histories while balancing user privacy and the utility of data consumer.

Although previous work has studied the trading problem of entire text-based Twitter data [27], the intrinsic feature of web

*The corresponding author

IWQoS '19, June 24–25, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6778-3/19/06…$15.00
https://doi.org/10.1145/3326285.3329060

browsing histories distinguishes ours from such work. In general, data contributors have diverse privacy preferences [12][17] with respect to various kinds of web browsing histories, which is defined as the maximum tolerant privacy loss on each category. For example, an ordinary data contributor usually cares more about health-related browsing records than other kinds like purchase or video records. Zhang *et al.* [27] only propose an outsourcing mechanism for social media data without consideration of Twitter users' privacy preferences. In addition, a practical trading mechanism should further take the data consumer' utility [11] into consideration. Unfortunately, existing work *e.g.*, [27] only measures the accuracy performance of the trading mechanism on evaluation experiments. Thus, it is still an unsolved problem to design a practical trading mechanism while considering data contributors' diverse privacy preferences on various kinds of web browsing histories and data consumers' realistic accuracy requirements simultaneously.

Designing a feasible trading strategy of various kinds of web browsing histories for practical data markets usually faces three major challenges. The first challenge is to satisfy data contributors' diverse privacy preferences for various kinds of web browsing histories. A concrete data trading format should be determined in consideration of any user's possible privacy leak while trading entire browsing histories directly. For insensitive categories of browsing histories, simple adoption of the popular Laplace mechanism [15][7] to perturb original records leads to large noise due to high dimensionality of user features. Nevertheless, no existing work solves the above two tough problems simultaneously. Thus, it is nontrivial to design a realistic trading strategy while complying with the maximum tolerant privacy loss for various kinds of browsing histories.

The second challenge is on achieving a suitable tradeoff between the data consumer's utility on the perturbed user-by-features matrix and data contributors' privacy protection. Stronger privacy protection usually implies more degraded utility, which is possibly inconsistent with the data consumer's accuracy requirement. Conversely, a higher utility may result in data contributors' unacceptable privacy leak due to violation of their privacy preferences. In addition, although previous work [5] also aims at addressing the utility-privacy tradeoff, they focus on the recommendation of personalized online service rather than the delivery of statistics results for the whole population. Therefore, it is still largely an open problem.

The last but not least challenge comes from preventing data contributors' possible strategic behaviors. In previous work, the data contributor's privacy valuation is supposed to be public knowledge [15][17], and they are compensated by a fixed private cost. This probably results in the data broker's selection bias under our assumption of unknown privacy valuation. Since the fixed compensation cost only attracts a group of data contributors whose privacy valuations are below this value, the corresponding user records cannot reflect the whole population and the result is probably biased. Therefore, a reasonable privacy compensation mechanism should adopt auction to compensate each data contributor. However, each participant possibly misreports the cost of his unit privacy loss for the higher benefit by auction. Thus, the truthfulness of the proposed privacy compensation mechanism further increases the complexity of designing a practical trading strategy.

In response to the challenges mentioned above, we propose *PEATSE*, as a new Privacy-prEserving dAta Trading framework for various kinds of web browSing historiEs, consisting of a data perturbation mechanism and a privacy compensation mechanism. For data perturbation, *PEATSE* exploits the popular user-by-features matrix as the ultimate trading format to generate the returned initial dataset. Due to high feature dimensionality of insensitive categories, *PEATSE* further adopts a modified Laplace mechanism to perturb the normalized data vector. For privacy compensation, *PEATSE* perturbs each chosen data contributor's data elements by the corresponding perturbation mechanism based on carefully picked Laplace noise, so as to achieve a desired tradeoff between the data consumer's utility and user privacy, while guaranteeing all desired economic properties simultaneously.

We highlight main contributions as follows.

- To the best of our knowledge, *PEATSE* is the first work that considers privacy-preserving data trading for various kinds of web browsing histories where a practical trading model is proposed.
- *PEATSE* satisfies data contributors' diverse privacy preferences on various kinds of web browsing histories while complying with the data consumer' accuracy requirement simultaneously.
- Through rigorous theoretical analysis and extensive evaluations, *PEATSE* guarantees data contributors' acceptable privacy loss once the data consumer's accuracy requirement is achieved, and thus achieves the desired tradeoff between user privacy and the data consumer's utility. Moreover, the evaluation results show the usefulness and feasibility of *PEATSE*, and that it achieves data contributors' individual rationality, truthfulness and budget balance.

2 PROBLEM FORMULATION

In this section, we introduce the system model, problem statement and preliminaries for practical data markets trading various kinds of web browsing histories.

2.1 System Model

As is illustrated in Fig.1, we consider a data trading market consisting of data contributors (*i.e.*, users), data consumers and a data broker, who is allowed to access users' web browsing histories as a trustworthy third-party platform like Infochimps [2].

2.1.1 Data Acquisition. Web browsing histories $\curlyvee \boldsymbol{W} = [\curlyvee \boldsymbol{W}^1, \ldots, \curlyvee \boldsymbol{W}^n]$ from N distinct data contributors, denoted as $N = \{1, 2, \ldots, n\}$, are formally given in Definition 1.

Definition 1. **(*Web Browsing History*)**. *For each data contributor $i \in N$, his web browsing history represents the set of n pairs of links and the page content,* i.e.*, $\curlyvee \boldsymbol{W}^i = \{(l_1, t_1), \ldots, (l_n, t_n)\}$, where l_i refers to the i-th visited URL, and t_i corresponds to the content of the web page.*

For the convenience of trading, the web browsing history should be converted to a structural format. A feasible solution is to divide $\curlyvee \boldsymbol{W}^i$ into K predefined categories like U = {Health, Finance, Social,

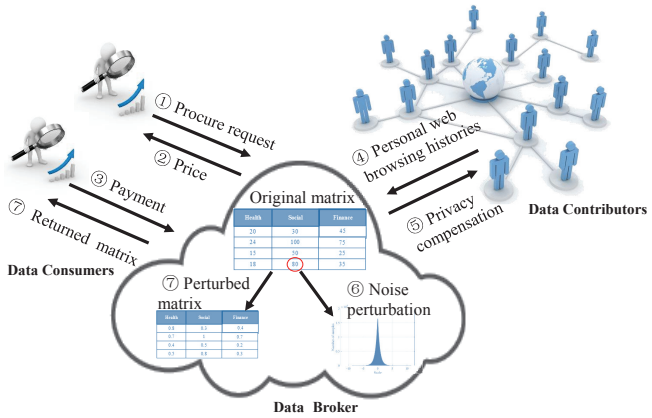

Figure 1: The system model for data markets trading personal web browsing histories.

Video, ...} in a coarse-grained way, where each pair (l_i, t_i) is mapped to the corresponding category according to the content t_i of the web page. The other categorization such as the extracted topic model [21] can be adopted to define the categories.

Given each data consumer's procurement request $\boldsymbol{Q}$, the data broker first locates the related data contributors' (*e.g.*, female users over 25 years old) browsing histories, and then generates the initial user-by-features matrix $\mathcal{D}_k$[1] with respect to each category k, and each matrix element means the user's total browsing times on some product. Next, our data model is given as follows.

Definition 2. ***(Data Vector)***. *Given the data consumer's interested user feature vector $\boldsymbol{o}^k = \{\boldsymbol{o}_1^k, \ldots, \boldsymbol{o}_d^k\}$ belonging to the matrix $\mathcal{D}_k$ (*e.g.*, the most frequently visited d e-commerce products), all chosen data contributors' data elements generate the corresponding data vector $\boldsymbol{x}_{(\boldsymbol{o}^k)} = \{\boldsymbol{x}_1^d, \ldots, \boldsymbol{x}_n^d\}$, where $\boldsymbol{x}_i^d \in R^d$ represents data contributor i's personal browsing records on d released user features.*

Therefore, we restrict attention to data contributors' numerical data [15]. For example, the data vector $\boldsymbol{x}_i^d$ may represent an individual's browsing frequency $\boldsymbol{x}_i^d = \{30, 20, \ldots\}$ on any feature vector $\boldsymbol{o}^k$ which includes features like $\boldsymbol{o}_1^k$ ='Amazon/electronics/iPhone'[2] and $\boldsymbol{o}_2^k$ ='Gallze/electronics/Samsung' belonging to the 'Social' category. Therefore, the data consumer intends to purchase the perturbed user-by-features matrix $\mathcal{D}_k^{'}$ on each category k, and further acquire the interested statistical results by aggregating data contributors' private information according to $\mathcal{D}_k^{'}$.

Data contributors usually have diverse privacy preferences[3] on various categories by the market survey [23]. The preference vector is denoted as $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K\}$, where $\boldsymbol{\phi}_k$ represents the maximum tolerant privacy loss on the category k to each data consumer. Given a threshold value $\overline{\boldsymbol{\phi}}$, the category k with less than $\overline{\boldsymbol{\phi}}$ privacy preference belongs to sensitive categories, otherwise insensitive one. Obviously, a smaller value $\boldsymbol{\phi}_k$ indicates this category is more sensitive to common data consumers, and needs stronger privacy protection while being leveraged to generate the returned matrix. Thus, more data perturbation should be added to the initial matrix $\mathcal{D}_k$ for sensitive categories than insensitive ones.

In addition to $\boldsymbol{\Phi}$, each data contributor submits the bid vector $\mathbf{c}_i = \{c_i^1, \ldots, c_i^K\}$, as his claimed cost of unit privacy loss (*i.e.*, private cost) for each category, which is unnecessarily equal to his real private cost $\overline{\boldsymbol{c}}_i = \{\overline{c}_i^1, \ldots, \overline{c}_i^K\}$ in terms of his possible strategic behaviors. Given each data contributor's actual privacy loss $\boldsymbol{\epsilon}_i = \{\epsilon_i^1, \ldots, \epsilon_i^K\}$ on each category, the data broker compensates him with $\xi_i(\boldsymbol{Q}) = \sum_{k=1}^K \xi_i^k(\boldsymbol{Q})$, where $\xi_i^k(\boldsymbol{Q})$ is his obtained compensation on the category k. Therefore, his utility is defined as follows:

$$u(\mathbf{c}_i) = \begin{cases} \sum_{k=1}^K (\xi_i^k(\boldsymbol{Q}) - \overline{c}_i^k * \epsilon_i^k) & if\ i \in G_k \\ 0 & otherwise \end{cases}, \quad (1)$$

where G_k is the set of data contributors whose browsing histories are leveraged to generate the fractional initial dataset $\mathcal{D}_k$, and the whole initial matrix is $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$.

2.1.2 Data Trading. The data broker is likely to trade the perturbed user-by-features matrix, rather than directly selling entire web browsing histories or single query answers[4] to data consumers. Since any feature vector within the returned matrix implies their interested data contributors' online activities, the corresponding statistical results can reflect preferences or tendencies of the user audience. For example, Nestle Corporation wonders to know the proportion of target users who ever purchased its product and properly care for posts so as to choose their preferences for Behavior Targeting on Twitter [4]. In our scenario, we assume data consumers only focus on the linear query [6] with respect to the returned matrix, which is given in Definition 3.

Definition 3. ***(Weighted Sum)***. *For each data vector $\boldsymbol{x}_{(\boldsymbol{o}^k)}$, the data consumer specifies the corresponding weight vector $\boldsymbol{w} = \{w_1, \ldots, w_n\}$, where w_i means that his importance or preference over data contributor i's elements $\boldsymbol{x}_i^d$. The real weighted sum is defined as $\zeta = \sum_{i \in N} w_i g(\boldsymbol{x}_i^d)$, where $\boldsymbol{x}_i^d$ represents the i^{th} row of the fractional initial matrix $\mathcal{D}_k$ and the function g maps this matrix row to $[0, 1]$.*

Each data consumer can acquire the corresponding weighted sum with respect to any interested feature vector on each category. Combined with the previous example, the data consumer possibly wonders to know the population of the users who purchase the iPhone but never buy the Samsung. In particular, the data consumer submits the procurement request $\boldsymbol{Q} = (q, \boldsymbol{\delta})$ to the data broker, where q specifies the group of target users, and the vector of error bound $\boldsymbol{\delta} = \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K\}$ represents his maximum tolerant variance for the statistical result on each category. Thus, the weighted sum for any feature vector $\boldsymbol{o}^k$ should belong to the acceptable range $[\zeta - \boldsymbol{\delta}_k, \zeta + \boldsymbol{\delta}_k]$. In addition, any data consumer's $\boldsymbol{\delta}$-accuracy requirement is given in Definition 4.

[1] also regarded as a real-valued matrix $\mathcal{D}_k \in \mathbb{R}^{p*d}$, where p and d represents the number of the related data contributors and the most frequently visited records in the whole dataset, respectively.

[2] The feature means target users ever visited webpages about iPhone for 30 times which belongs to electronics on Amazon, and the classified titles like 'Amazon', 'electronics' can be extracted by the preprocessing of page content t_i.

[3] also called private budget, and is supposed to be public knowledge.

[4] The reason why the data broker sells the perturbed matrix rather than the single statistical result is that the private perturbed database can keep a data consumer from revealing his real analysis task to the data broker, and even exhausting the private budget before acquiring the interested query answer.

Definition 4. **($\boldsymbol{\delta}$-accuracy requirement)**. *For each category k, the data consumer's any weighted sum query from the returned perturbed matrix $\mathcal{D}_k^{'}$ satisfies the $\boldsymbol{\delta}_k$-accuracy requirement if and only if the output y after the data perturbation mechanism satisfies:*

$$Pr(|y - \zeta| \geq \boldsymbol{\delta}_k) \leq \rho, \tag{2}$$

where the above inequation holds for any possible output y. Note that the probability value ρ can be any constant less than $\frac{1}{\sqrt{e}+1}$, e.g., $\rho = \frac{1}{3}$.

It is worth noting that the data consumer's utility is guaranteed when the traded user-by-features matrix satisfies the above $\boldsymbol{\delta}_k$-accuracy requirement. Without loss of generality, the data consumer has a higher utility when the query answer belongs to the acceptable range with the higher probability $1 - \rho$ and the deviation value from the true value is smaller.

Given the request $Q = (q, \boldsymbol{\delta})$, the data broker charges the data consumer with $\psi(Q) = \sum_{k=1}^{K} \psi^k(Q)$, where $\psi^k(Q)$ corresponds to the charged price for any category k. Obviously, if the error bound δ_k for any feature vector on the category k is smaller, the charged price $\psi^k(Q)$ would be larger. Moreover, any related data contributor's privacy break ϵ_i^k should be higher, and thus the compensation $\xi_i^k(Q)$ is also larger. Without loss of generality, the proposed trading mechanism is balanced if and only if the charged price for the data consumer can cover the total compensation cost paid for all the related data contributors, *i.e.*, $\psi(Q) - \sum_{i \in G} \xi_i(Q) \geq 0$, where $G = \{G_1, \ldots, G_K\}$.

2.2 Problem Statement and Preliminaries

We first introduce the attack model as a preliminary, and then present our design objectives.

2.2.1 Attack Model. The delivery of even anonymous web browsing histories probably causes the user-linkage attack [26], which can be carried out by a curious or even malicious data consumer. The attacker can locate the target user according to the prior knowledge about this user's open browsing history on social media platform such as the recently posted tweets on Twitter. Consequently, the attacker acquires this target user's other web browsing histories, especially including health-related sensitive records.

For our scenario, although only browsing times for each category are released in the returned matrix, unperturbed counts still lead to the same attack. For example, the malicious attacker can infer the target user's position within the received matrix when he counts the number of this victim's tweets on Twitter during given periods, matches the number with candidates' records, and finally obtains this victim's other sensitive information.

2.2.2 Problem Statement. We aim at designing a practical trading strategy with respect to the aforementioned general trading problem for various kinds of web browsing histories. After receiving the data consumer's request, the data broker first initializes the user-by-features matrix $\mathcal{D}$ through all related data contributors, including the target users' various kinds of browsing histories and the corresponding browsing times within the specified period of request. In addition, the anonymous user ID is assigned to each related data contributor in terms of possible privacy break. Next, the data broker converts the initial matrix $\mathcal{D}$ to the perturbed matrix $\mathcal{D}^{'}$, where users' browsing times on each category k are added Laplace noise so as to defend against the malicious data consumer's possible inference. Moreover, it is essential to satisfy the following requirements simultaneously while designing the trading strategy.

- *Diverse Privacy Preservation*: For each category k, the attacker can match any anonymous user ID with the real user by browsing times in the delivered matrix $\mathcal{D}_k^{'}$ with the insignificant probability, where $\mathcal{D}_k^{'}$ is the perturbed user-by-features matrix related to the category k.
- *Accuracy Requirement*: The data consumer can achieve $\boldsymbol{\delta}_k$-accuracy requirement for any weighted sum query with respect to each category k from $\mathcal{D}_k^{'}$.
- *Budget Balance*: The charged price for any data consumer can be distributed to all chosen data contributors in an affordable manner.
- *Incentive Compatibility*: For any data contributor i, he can achieve a higher utility bidding for the truthful bid vector $\mathbf{c}_i$ than the untruthful bid $\tilde{\mathbf{c}}_i$, *i.e.*, $u(\mathbf{c}_i, \mathbf{c}_{-i}) \geq u(\tilde{\mathbf{c}}_i, \mathbf{c}_{-i})$, where $\mathbf{c}_{-i}$ represents the set of bid vectors except $\mathbf{c}_i$.
- *Individual Rationality*: Any data contributor i has nonnegative utility for the truthful bid vector, *i.e.*, $u(\mathbf{c}_i) \geq 0$.

3 PRIVACY-PRESERVING WEB BROWSING HISTORY TRADING

In this section, we present *PEATSE* with the aforementioned design goals. *PEATSE* is elaborated as the following steps. *First*, we describe the formulation method of the initial user-by-features matrix. *Second*, according to the generated initial matrix, we propose the data perturbation component of *PEATSE* to satisfy data contributors' diverse privacy preferences for each category. *Finally*, with respect to the perturbed matrix, we introduce the privacy compensation component of *PEATSE* to quantify each data contributor's privacy loss, and calculate the related compensation cost.

3.1 Generate Initial Matrix

After receiving the data consumer's request $Q = (q, \boldsymbol{\delta})$, the data broker first finds the related data contributors' various kinds of browsing histories, and generates an original matrix. Each record is represented as an event $e_i = (\tau_i, k_i, t_i)$, where τ_i means browsing times, k_i denotes the category this record belongs to, and t_i is the content of the web page. And then the broker maps each record to the corresponding category by the page content t_i.

3.1.1 Initial Matrix. Next, the data broker generates the detailed user-by-features matrix. For insensitive categories like 'Social', the broker first extracts the key user feature of all related data contributors' any event e_i according to its content t_i, then counts the frequency of each user feature as a term, and finally chooses the most frequent d features to be added to the returned matrix $\mathcal{D}_k$. It is reasonable to choose the features most related data contributors have, because infrequent features are more likely to expose those data contributor's real identity. Moreover, for sensitive categories like 'Health', we remove the most user features in consideration of data contributors' high privacy concern, and only retain the rough user feature.

Note that not all data contributors' browsing information would be delivered in the returned matrix because of the data consumer's acceptable error bound. Therefore, some related data contributors' browsing times are possibly unavailable once the data consumer's required accuracy guarantee is achieved. We will discuss it in Section 3.C. Consequently, the initial matrix can be taken as an example in table 1, where N/A means the related records are not released. In table 1, 4 data contributors' browsing histories are released with 3 categories, where 'Health' belongs to sensitive categories, and thus the feature vector only includes one feature 'Surgery/Fracture', but other insensitive categeory like 'Social' has more released user features.

3.1.2 Normalization. Moreover, data contributor i's element x_{ij} on any user feature j in the initial matrix is further normalized to the range $[0, 1]$ because most data contributors usually have diverse browsing times for various user features.

3.2 Perturb Matrix

First, we present the first component of *PEATSE*, namely the data perturbation component for the returned matrix. Even the normalized initial matrix probably leads to the user-linkage attack after the denormalization with the information of some data elements for each user feature, which assists the attacker to infer the victim's real identity until he links the corresponding records to a real user. As a result, to circumvent data contributors' identity exposure with a high possibility, the data broker perturbs the initial matrix in different ways for various kinds of browsing histories.

Existing work adopt the popular differential privacy [7][8] to perturb the statistical query answer, as an effective approach first proposed by Dwork *et al.* [7]. Recall that the main idea of the Laplace mechanism is to add Laplace noise for the returned statistical result, so as to limit the attacker's possibility of inferring some individual's private information. However, for our scenario, we return the user-by-features matrix rather than a perturbed query answer. Hence, we perturb the data element of the returned matrix.

Since the data consumer can usually acquire multiple weighted sum queries from the purchased matrix, a new general notion Φ-indistinguishability is required in order to handle these queries simultaneously. Given the data vector $\boldsymbol{x}_{(\boldsymbol{o}^k)} = \{\boldsymbol{x}_1^d, \ldots, \boldsymbol{x}_n^d\}$ for any feature vector $\boldsymbol{o}^k$ on each category k, where $\boldsymbol{x}_i^d$ corresponds to data elements of the d-dimensional feature vector belonging to data contributor i. Let $\boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)}$ denote the neighboring data vector without data contributor i's records $\boldsymbol{x}_i^d$. The formal definition of Φ-indistinguishability is given in Definition 5.

Definition 5. **(Φ-*indistinguishability*)**. *A randomized mechanism $\mathcal{M}_f$ satisfies Φ-indistinguishability if any two neighboring vectors $\boldsymbol{x}_{(\boldsymbol{o}^k)}, \boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)} \in R^{n*d}$ related to the feature vector $\boldsymbol{o}^k$ belongs to the category k, and for any possible output S, we have*

$$\frac{Pr(\mathcal{M}_f(\boldsymbol{x}_{(\boldsymbol{o}^k)}) = S)}{Pr(\mathcal{M}_f(\boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)}) = S)} \leq e^{\boldsymbol{\phi}_k}, \tag{3}$$

where $Pr(\mathcal{M}_f(\boldsymbol{x}_{(\boldsymbol{o}^k)}) = S)$ means the probability density of the random variable after the general function $\mathcal{M}_f$ on $\boldsymbol{x}_{(\boldsymbol{o}^k)}$, and $\boldsymbol{\phi}_k$ is the privacy budget for any category k. The larger value $\boldsymbol{\phi}_k$ indicates a larger privacy loss because of the higher output difference, and the attacker is more likely to find the change of the initial matrix. Specifically, the numeric vector-valued function f maps the matrix to vectors of reals, and then the sensitivity of the function f over the data vector $\boldsymbol{x}_{(\boldsymbol{o}^k)}$ is defined as the maximum difference between $f(\boldsymbol{x}_{(\boldsymbol{o}^k)})$ and $f(\boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)})$, *i.e.*,

$$\Delta f_i = \max_{\boldsymbol{x}_{(\boldsymbol{o}^k)}, \boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)}} ||f(\boldsymbol{x}_{(\boldsymbol{o}^k)}) - f(\boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)})||_1, \tag{4}$$

where $\boldsymbol{x}_{(\boldsymbol{o}^k)} \in R^{n \times d}$.

Next, we elaborate the data perturbation component $\mathcal{M}_f$ of *PEATSE* for each category. In terms of the data consumer's acceptable error bound $\boldsymbol{\delta}$, the data broker only chooses a proportion of the related data contributors so as to deliver their perturbed data elements. Because fractional data contributors' records can generate the query answer satisfying the data consumer's $\boldsymbol{\delta}$−accuracy requirement in Definition 4, the data broker is willing to pay for eligible data perturbation with the less compensation cost. Specifically, we first introduce two diverse perturbation mechanisms for data elements from different categories, and then propose the privacy compensation component of *PEATSE* to choose the proportion of data contributors in order to satisfy the data consumer's accuracy requirement for a realistic scenario in Section 3.C..

3.2.1 Sensitive Categories. For sensitive categories like U_1 = 'Health', we add Laplace noise[5] to chosen data contributors' data elements, *i.e.*,

$$\mathcal{M}_f(\beta) = \beta + Lap(\Delta f_i / \boldsymbol{\phi}_k), \tag{5}$$

where β refers to any chosen data contributor' data elements belonging to the category U_1, and dimensionality of the data vector $\boldsymbol{x}_{(\boldsymbol{o}^k)}$ is 1, *i.e.*, $\boldsymbol{x}_{(\boldsymbol{o}^k)} \in R^{n \times 1}$ in terms of the only delivered user feature. And added noise follows from the one-dimensional Laplace distribution $Lap(\lambda)$, with the density function $h(x) = e^{(-|x|/\lambda)}$ where the mean is 0 and the variance is $2\lambda^2$. Consequently, added noise satisfies the equation (6).

$$Lap(\Delta f_i / \boldsymbol{\phi}_k) \propto e^{-\boldsymbol{\phi}_k |x| / \Delta f_i}, \tag{6}$$

where $\Delta f_i = |w_i|(\overline{\beta}_i - \underline{\beta}_i)$, and $\overline{\beta}_i$ and $\underline{\beta}_i$ represent the upper bound and the lower bound of any data element $\boldsymbol{x}_i^d$, respectively.

Table 1: A toy example for the initial matrix with 3 categories in 2 days

Health	*Social*		*Finance*	
Surgery/ Fracture	*Amazon/ electronics/ iPhone XP*	*Gallze/ electronics/ Samsung*	*Stocks/ Profile*	*Funds/ Profile*
2	*30*	*20*	*N/A*	*10*
3	*25*	*15*	*8*	*9*
4	*N/A*	*21*	*14*	*16*
1	*28*	*18*	*10*	*8*

[5]The reason why the data broker chooses the Laplace distribution rather than the other distributions like the Uniform or Gaussian distribution is that only Laplace noise matches up perfectly with the definition of differential privacy.

Theorem 1. *The proposed perturbation mechanism $\mathcal{M}_f$ for sensitive categories satisfies $\boldsymbol{\phi}_k$-indistinguishability.*

Proof. In consideration of the probability density function $h(x) = e^{(-|x|/\lambda)}$, then for any two data vectors $z, z^{'} \in \boldsymbol{x}_{(\boldsymbol{o}^k)}$, we have $\frac{h(z)}{h(z^{'})} = e^{\boldsymbol{\phi}_k(|z^{'}|-|z|)}$ combined with added noise and the sensitivity 1 of the function f after the normalization of the initial matrix. Moreover, for any two data vectors $\boldsymbol{x}_{(\boldsymbol{o}^k)}, \boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)} \in R^{n\times 1}$, the corresponding numeric functions $f(\boldsymbol{x}_{(\boldsymbol{o}^k)})$ and $f(\boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)})$ only have one gap. Thus, for any possible output $S \in R$, we have

$$\begin{aligned}\frac{Pr(\mathcal{M}_f(\boldsymbol{x}_{(\boldsymbol{o}^k)}) = S)}{Pr(\mathcal{M}_f(\boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)}) = S)} &= \frac{h(S - f(\boldsymbol{x}_{(\boldsymbol{o}^k)}))}{h(S - f(\boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)}))} \\ &= e^{\boldsymbol{\phi}_k(|f(\boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)})|-|f(\boldsymbol{x}_{(\boldsymbol{o}^k)})|)} \\ &\le e^{\boldsymbol{\phi}_k(|f(\boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)})-f(\boldsymbol{x}_{(\boldsymbol{o}^k)})|)} \le e^{\boldsymbol{\phi}_k}\end{aligned} \tag{7}$$

The first line comes from the definition of added noise, and the third line follows from the triangle inequality. Due to the sensitivity 1 of the function f again, the final line holds. Thus, the theorem follows. □

3.2.2 Insensitive Categories. For insensitive categories like U_3 = 'Social', multiple user features as the feature vector are returned in the matrix, indicating feature dimensionality d of this category is usually high (*i.e.*, $\boldsymbol{x}_{(\boldsymbol{o}^k)} \in R^{n\times d}$), shown in table 1. A natural approach is to perturb each dimensionality of this category with corresponding Laplace noise, *i.e.*,

$$\mathcal{M}_f(V_i) = V_i + (Y_{i1}, Y_{i2}, \ldots, Y_{id}), \tag{8}$$

where $V_i \in R^d$ is the data vector of the feature vector belonging to the same category, and Y_{ij} are drawn i.i.d from the equation (6).

However, since the change of any user feature on this category increases the sensitivity of the function f, the prior method leads to a larger deviation of Laplace noise with the higher dimensionality of the feature vector according to the equation (6). Specifically, if each data element for the feature vector of U_3 is introduced noise by the equation (5), then added noise for the higher dimensionality becomes non-negligible compared with the norm (*i.e.*, 1) of each data element after the normalization. Consequently, the data consumer's utility of the returned matrix for this category would decrease largely due to the large deviation of noise. Therefore, the traditional Laplace mechanism fails.

Inspired by previous work [27], we add noise to the data vector for this category in each row, *i.e.*,

$$\mathcal{M}_f(V_i) = V_i + \omega\mathcal{I}, \tag{9}$$

where ω represents the random distance between the original vector V_i and the perturbed vector $V_i^{'}$, and $\mathcal{I}$ refers to a d-dimensional unit random vector. Note that the new perturbation mechanism requires distance noise drawn from the Laplace distribution $Lap(\gamma_{max}\Delta f_i/\boldsymbol{\phi}_k)$, where γ_{max} represents the maximum distance between any two data vector V_i and V_j, and $\gamma_{max} \le \sqrt{d}$ because each data element has the maximum value 1 after the normalization. Note that for insensitive categories, $\Delta f_i = |w_i|||\overline{V}_i - \underline{V}_i||_1$, where $\overline{V}_i$ and $\underline{V}_i$ is the supremum and infimum of the data vector V_i, respectively.

Theorem 2. *For insensitive categories, the randomized mechanism $\mathcal{M}_f$ also satisfies $\boldsymbol{\phi}_k$-indistinguishability.*

Proof. According to the new perturbation mechanism in equation (9), we have

$$\begin{aligned}\frac{Pr(\mathcal{M}_f(\boldsymbol{x}_{(\boldsymbol{o}^k)}) = S)}{Pr(\mathcal{M}_f(\boldsymbol{x}^{(i)}_{(\boldsymbol{o}^k)}) = S)} &= \frac{h(\omega(\overline{V}, V_i)\mathcal{I}_1)}{h(\omega(\overline{V}, V_j)\mathcal{I}_2)} \\ &= e^{\frac{\boldsymbol{\phi}_k}{\gamma_{max}}(\omega(\overline{V}, V_i)-\omega(\overline{V}, V_j))} \\ &\le e^{\frac{\boldsymbol{\phi}_k}{\gamma_{max}}\omega(V_i, V_j)} \le e^{\boldsymbol{\phi}_k}\end{aligned} \tag{10}$$

The first line is based on the definition of the perturbation mechanism, and the second line comes from the fact that any two d-dimensional unit vector are generated with the same probability, and the third line holds due to the triangle inequality. □

3.3 Privacy Compensation

In this section, we further consider the second component of *PEATSE*, namely the privacy compensation component for data contributors whose various kinds of browsing histories are leveraged to generate the returned perturbed matrix. Because the data consumer purchases the perturbed matrix so as to acquire weighted sum queries with the acceptable error bound in Definition 4 for any feature vector on each category, data contributors' private information like released data elements $\boldsymbol{x}_i^d$ has to be leaked partly by answering each query, and thus they must be compensated. Next, we propose a practical privacy compensation mechanism for a realistic data trading scenario.

3.3.1 A realistic scenario. Suppose that the data broker's budget[6] is B for any returned user-by-features matrix. Since fixed privacy compensation probably leads to the biased statistical answer, each data contributor usually reports his private cost $\mathbf{c}_i = \{c_i^1, \ldots, c_i^K\}$ by auction in a realistic scenario. Moreover, considering that data contributors have diverse privacy budgets for different categories, it is inefficient to achieve the same lowest privacy guarantee on all categories. Thus, it is more reasonable to quantify any user's privacy loss and calculate privacy compensation separately with respect to each category.

Next, we give an intuitive privacy compensation mechanism in Algorithm 1, which demonstrates a data broker has to purchase the amount of privacy from certain data contributors to achieve the data consumer's desired accuracy requirement. Suppose that each data contributor's upper bound of the privacy loss on each category is given as the input, *i.e.*, $\overline{\boldsymbol{\epsilon}}_i = \{\overline{\epsilon}_i^1, \ldots, \overline{\epsilon}_i^K\}$, which will be further discussed in Sec 3.C.(3).

In consideration of the limited budget B, the data broker picks data contributors from those with the less than $\frac{B}{K*r}$ of compensation cost in advance according to line 5. Next, the broker calculates the number of data contributors who are willing to accept the minimum privacy loss $\frac{1}{4\boldsymbol{\delta}_k}$ according to line 6-11.

[6]The budget is assumed to be given when the sale price of the returned matrix is set as the optimal price based on data consumers' bid profiles by Bayesian optimal auction [28].

Algorithm 1: Privacy Compensation Mechanism

Input: Initial matrix $\mathcal{D}$, the upper bound set of privacy loss $\overline{\boldsymbol{\epsilon}} = \{\overline{\boldsymbol{\epsilon}}_1, \ldots, \overline{\boldsymbol{\epsilon}}_n\}$, the data consumer's error bound $\boldsymbol{\delta}$, the maximum distance γ_{max}, the set of data contributors' private cost $\mathbf{c}$, and the budget B.

Output: A set $G = \{G_1, \ldots, G_K\}$ of selected data contributors, the payment vector $\mathbf{p}$ and the perturbed matrix $\mathcal{D}'$.

1 $\{t_1, \ldots, t_K\} \leftarrow \mathbf{0}$, $G \leftarrow \vec{\emptyset}$, $\mathbf{p} \leftarrow \mathbf{0}$;
2 // *Data Contributor Selection* ;
3 **for** $k = 1$ *to* K **do**
4 Sort all data contributors from $\mathcal{D}$ in the increasing order of c_i^k;
5 Find the largest index r such that $c_r^k * \frac{1}{4\boldsymbol{\delta}_k} \leq \frac{B}{K*r}$;
6 **for** $i = 1$ *to* r **do**
7 $\epsilon_i^k = \frac{1}{4\boldsymbol{\delta}_k}$;
8 **if** $\epsilon_i^k \leq \overline{\epsilon}_i^k$ **then**
9 $t_k = t_k + 1$;
10 **end**
11 **end**
12 **if** $t_k \geq |N| - 4\boldsymbol{\delta}_k$ **then**
13 Choose the first $|G_k| = |N| - 4\boldsymbol{\delta}_k$ data contributors as winners;
14 **for** $i = 1$ *to* $|G_k|$ **do**
15 $G_k \leftarrow G_k \bigcup \{i\}$;
16 $\mathcal{M}_f(\beta) = \beta + Lap(\Delta f_i / \epsilon_i^k)$ to $\mathcal{D}_k'$;
17 Or $\mathcal{M}_f(V_i) = V_i + Lap(\gamma_{max}\Delta f_i / \epsilon_i^k)\mathcal{I}$ to $\mathcal{D}_k'$;
18 // *Payment Scheme* ;
19 $p_i = p_i + \min(\frac{B}{K*r}, c_{r+1}^k * \frac{1}{4\boldsymbol{\delta}_k})$;
20 **end**
21 **end**
22 **end**
23 **return** $(G, \boldsymbol{p}, \mathcal{D}')$;

By line 12-20, if the number is larger than $|N| - 4\boldsymbol{\delta}_k$, then the broker chooses the first $|G_k|$ data contributors, and perturbs each winner's data elements with actual privacy loss of $\frac{1}{4\boldsymbol{\delta}_k}$ based on the corresponding category of the data perturbation mechanism. Note that any unselected data contributors' data elements would not be released in the returned matrix $\mathcal{D}'$. Finally, in line 19, we distribute the payment $\xi_i^k(\boldsymbol{Q}) = \min(\frac{B}{K*r}, c_{r+1}^k * \frac{1}{4\boldsymbol{\delta}_k})$ to each winning data contributor i on each category k. Note that for our scenario, each data consumer is charged for $\psi^k(\boldsymbol{Q}) = \frac{B}{K}$ on each category k. For the page limit, the time complexity of Algorithm 1 is given as $O(K \cdot Nlog(N))$.

3.3.2 Theoretical Analysis. Algorithm 1 shows that the data broker can satisfy the data consumer's accuracy requirement in Definition 4 once he purchase the ϵ_i^k amount of data privacy from at least $|G_k|$ data contributors on each category, where ϵ_i^k and $|G_k|$ are only related to the data consumer's error bound $\boldsymbol{\delta}_k$. Next, we demonstrate the efficiency of Algorithm 1 in Theorem 3.

Theorem 3. *The data consumer can achieve desired $\boldsymbol{\delta}$-accuracy guarantee if $\mathcal{M}_f$ meets the following conditions on each category k: 1) There are at least $|G_k|$ data contributors whose privacy loss is larger than $\frac{1}{4\boldsymbol{\delta}_k}$,* i.e., *$\epsilon_i^k \geq \frac{1}{4\boldsymbol{\delta}_k}$, for $i \in G_k$; 2) $|G_k| \geq N - 4\boldsymbol{\delta}_k$.*

Proof. For any category k, suppose that the mechanism $\mathcal{M}_f$ satisfies $\boldsymbol{\delta}_k$-accuracy, and the set of selected data contributors in the returned matrix $\mathcal{D}_k'$ are G_k and $\epsilon_i^k \geq \frac{1}{4\boldsymbol{\delta}_k}, i \in G_k$. First assume that the opposite of condition 2) holds, *i.e.*, $|G_k| < N - 4\boldsymbol{\delta}_k$, then we derive the false fact so as to prove the pseudo-proposition of our hypothesis.

Let $\overline{G}_k$ denote the set of unselected data contributors, *i.e.*, $\overline{G}_k = N \backslash G_k$, and further $|\overline{G}_k| > 4\boldsymbol{\delta}_k$. Recall that the data consumer obtains the weighted sum query for each category. Let S denote data consumers' acceptable output set satisfying $\boldsymbol{\delta}_k$-accuracy guarantee, *i.e.*, $S = \{y \in \mathbb{R}^7 | y - \zeta| < \boldsymbol{\delta}_k\}$, where $\zeta = \sum_{i \in \mathcal{D}_k} w_i g(\boldsymbol{x}_i^d)$. By the definition 4, we have any output y by the mechanism $\mathcal{M}_f$ satisfies $Pr(y \in S) \geq 1 - \rho$.

The set $\overline{G}_k$ is further divided into two parts, *i.e.*, $\overline{G}_k^1 = \{i \in \overline{G}_k | g(\boldsymbol{x}_i^d) = 1\}$ and $\overline{G}_k^0 = \{i \in \overline{G}_k | g(\boldsymbol{x}_i^d) = 0\}$. And obviously we have $\max(|\overline{G}_k^0|, |\overline{G}_k^1|) > 2\boldsymbol{\delta}_k$. Suppose that $|\overline{G}_k^1| > 2\boldsymbol{\delta}_k$ (the other case is the same). Next, we consider the other distant matrix $\mathcal{D}_k^L$ which has hamming distance $|L|$ with $\mathcal{D}_k$, where L denotes the set of different indices between $\mathcal{D}_k^L$ and $\mathcal{D}_k$, and $L \subset \overline{G}_k^1, |L| = 2\boldsymbol{\delta}_k$. Therefore, for any subscript $i \in L$, $g'(\boldsymbol{x}_i^d) = 0$, whereas $g'(\boldsymbol{x}_i^d)$ remains the same.

Next, we compare the difference between the probability of outputs by the distance matrix $\mathcal{D}_k^L$ and the initial matrix $\mathcal{D}_k$ separately belonging to the acceptable output set S. Let $\mathcal{D}_k^i$ and $\mathcal{D}_k^{i+1}$ represent two neighboring matrixs which only differ in the i^{th} indices of L, for any $0 < i < |L|$, then we have:

$$\begin{aligned} \frac{Pr(\mathcal{M}_f(\mathcal{D}_k) = y)}{Pr(\mathcal{M}_f(\mathcal{D}_k^L) = y)} &= \prod_{i=1}^{|L|-1} \frac{Pr(\mathcal{M}_f(\mathcal{D}_k^i) = y)}{Pr(\mathcal{M}_f(\mathcal{D}_k^{i+1}) = y)} \\ &\leq \prod_{i \in L} e^{\epsilon_i^k} \\ &= e^{\sum_{i \in L} \epsilon_i^k}, \end{aligned} \tag{11}$$

where the second line holds once selected data contributors' data elements are perturbed based on privacy loss ϵ_i^k, which is also smaller than the upper bound $\overline{\epsilon}_i^k$. Therefore, *PEATSE* naturally achieves $\boldsymbol{\phi}_k$-indistinguishability. Moreover, for any output $\hat{y}$, as the weighted sum based on the distance matrix $\mathcal{D}_k^L$ by the mechanism $\mathcal{M}_f$, we have:

$$\begin{aligned} Pr(\hat{y} \in S) &\geq e^{-\sum_{i \in L} \epsilon_i^k} Pr(y \in S) \\ &\geq e^{(-\frac{1}{4\boldsymbol{\delta}_k} \times 2\boldsymbol{\delta}_k)} \times (1 - \rho) \\ &= \frac{1 - \rho}{\sqrt{e}} \end{aligned} \tag{12}$$

Obviously, when $\rho < \frac{1}{\sqrt{e}+1}$, we have $\frac{1-\rho}{\sqrt{e}} > \rho$. Note that $y^* \geq \zeta + 2\boldsymbol{\delta}_k$, where y^* is the weighted sum over all the perturbed data vector, *i.e.*, $y^* = \sum_{i \in \mathcal{D}_k^L} w_i g'(\boldsymbol{x}_i^d)$. We have $|\hat{y} - \zeta| < \boldsymbol{\delta}_k$ by the definition if $\hat{y} \in S$. Moreover, we further have $|\hat{y} - y^*| = |(y^* - \zeta) - (\hat{y} - \zeta)| \geq |y^* - \zeta| - |\hat{y} - \zeta| \geq \boldsymbol{\delta}_k$ according to the triangle inequality, and the inequality holds with the probability of larger than ρ, which deviates from the fact of Definition 4. Therefore, our original hypothesis fails, and we prove $|G| \geq N - 4\boldsymbol{\delta}_k$. □

[7] $\mathbb{R}$ is the set of real numbers.

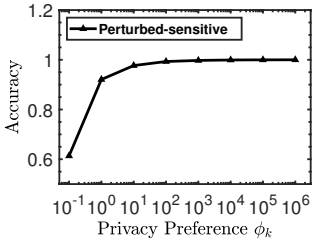

Figure 2: Privacy Preference ϕ_k vs. Accuracy for sensitive categories.

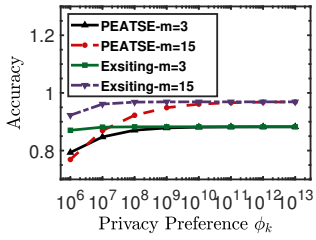

Figure 3: Privacy Preference ϕ_k vs. Accuracy for insensitive categories.

Based on theorem 3, *PEATSE* indeed achieves the desirable tradeoff by setting a fixed privacy loss for chosen data contributors on condition that it satisfies the data consumer's $\boldsymbol{\delta}$-accuracy guarantee. Finally, we show economic properties of *PEATSE* in theorem 4.

Theorem 4. *The proposed practical privacy compensation component of* PEATSE *achieves individual rationality, truthfulness as well as budget balance.*

Proof. First, we prove *PEATSE* is truthful. According to the work [18], the mechanism is truthful if and only if the data contributor selection algorithm is monotonic, and each winning data contributor is paid the critical payment.

First, we prove that any data contributor cannot increase his utility by misreporting his private cost regardless of others' report. Suppose that any chosen data contributor i on any category k reports a lower private cost $\tilde{c}_i^k \leq c_i^k$. According to the Algorithm 1, his claimed cost is still lower than c_{r+1}^k, and he still wins in terms of unchanged privacy loss. Thus, the monotonicity is satisfied.

Next, we prove the chosen data contributor is paid the critical payment. Suppose if data contributor i reports a higher cost than c_{r+1}^k, *i.e.*, $c_i^k \geq c_{r+1}^k$, i will lose and another data contributor with lower than or equal to c_{r+1}^k will replace him as the new winner. Thus, his utility decreases to 0. Conversely, this data contributor still wins but is yet paid $c_{r+1}^k * \epsilon_i^k$, and his utility is never improved because of the misreporting. Therefore, *PEATSE* satisfies truthfulness.

For unselected data contributors, their privacy loss on this category k is 0, and his compensation $\xi_i^k(\boldsymbol{Q}) = 0$. For the set G_k of chosen data contributors, each data contributor's privacy loss on each category k is $\epsilon_i^k = \frac{1}{4\boldsymbol{\delta}_k}$. The compensation is $\xi_i^k(\boldsymbol{Q}) = \min(\frac{B}{K*r}, c_{r+1}^k * \epsilon_i^k) \geq \bar{c}_i^k * \epsilon_i^k = c_i^k * \epsilon_i^k$ because of $c_i^k \leq c_{r+1}^k$ according to the sorting operation in Algorithm 1, and $c_i^k * \epsilon_i^k \leq \frac{B}{K*r}$ due to line 5. Thus, each data contributor is individual rational.

For each category k, all data contributors' total compensation $\sum_{i \in G_k} \xi_i^k(\boldsymbol{Q}) = |G_k| * \min(\frac{B}{K*r}, c_{r+1}^k * \epsilon_i^k) \leq r * \min(\frac{B}{K*r}, c_{r+1}^k * \epsilon_i^k) \leq \frac{B}{K}$, and *PEATSE* satisfies budget balance. □

3.3.3 Upper Bound of Privacy Loss. We further give the upper bound of any data contributor's privacy loss. First, we formally define the data contributor's privacy loss on each data vector $\boldsymbol{x}_i^d$, which is regarded as the bound of all the possible values according

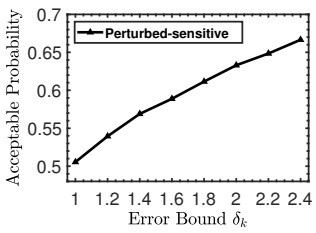

Figure 4: Acceptable Probability vs. Error Bound δ_k for sensitive categories.

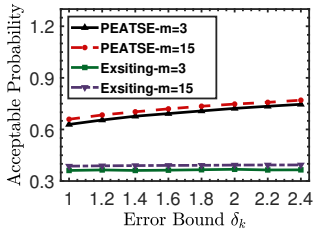

Figure 5: Acceptable Probability vs. Error Bound δ_k for insensitive categories.

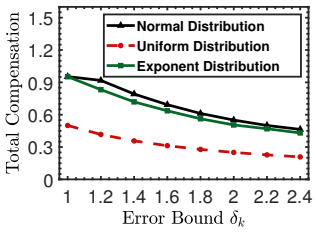

Figure 6: Total compensation vs. Error Bound δ_k

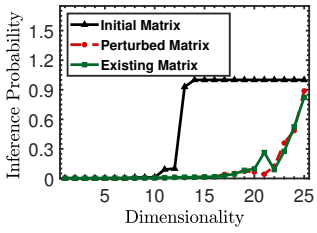

Figure 7: Inference Probability vs. Dimensionality d.

to any two neighboring vectors (*i.e.*, $\boldsymbol{x}_{(\boldsymbol{o}^k)}$ and $\boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)}$) with and without participation of his elements $\boldsymbol{x}_i^d$.

Definition 6. ***(Privacy Loss)***. *If the proposed perturbation mechanism $\mathcal{M}_f$ generates the perturbed matrix $\mathcal{D}'$, then data contributor i's privacy loss over the data element $\boldsymbol{x}_i^d$ is defined as:*

$$\epsilon_i^k(\mathcal{M}) = sup_{\boldsymbol{x}_{(\boldsymbol{o}^k)}, S} \left| log \frac{Pr(\mathcal{M}(\boldsymbol{x}_{(\boldsymbol{o}^k)}) = S)}{Pr(\mathcal{M}(\boldsymbol{x}_{(\boldsymbol{o}^k)}^{(i)}) = S)} \right|. \tag{13}$$

Next, we derive the upper bound $\bar{\epsilon}_i^k$ of each data contributor's privacy loss $\epsilon_i^k(\mathcal{M})$ on each category k, which depends on the data consumer's acceptable error bound $\boldsymbol{\delta}_k$ and the sensitivity of the function f on the data vector $\boldsymbol{x}_i^d$, shown in Theorem 5.

Theorem 5. *Each data contributor's privacy loss on each category k is bounded by $\epsilon_i^k(\mathcal{M}) \leq \frac{\Delta f_i}{\sqrt{\boldsymbol{\delta}_k/2}}$.*

Proof. For the page limit, we omit the detailed proof. Consequently, in Algorithm 1, the data contributor who has the sensitivity $\Delta f_i = 1$ would be chosen as the winner when the upper bound is larger than $\frac{1}{4\boldsymbol{\delta}_k}$, *i.e.*, $\boldsymbol{\delta}_k \geq \frac{1}{32}$. □

4 EVALUATION

In this section, we evaluate the performance of *PEATSE* in consideration of the data consumer's accuracy guarantee and data contributors' privacy protection.

4.1 Evaluation Settings

4.1.1 Dataset. We first introduce a real-world dataset collected by some handset manufacturer, which includes 12473 web users'

various kinds of web browsing histories on February 2009, ranging from E-commerce purchasing records, visited finance websites, to video websites. We count each data contributor's browsing times on each user feature for any category, and then generate the initial matrix. For example, for the 'Social' category, if some data contributor's page content is related to the extracted user feature 'Amazon/electronics/iPhone', then we increase his browsing times.

4.1.2 Settings. We generate data contributors' unit private costs according to three distributions, *i.e.*, normal distribution, uniform distribution and exponential distribution. For any category k, suppose that data contributors' privacy preference $\boldsymbol{\phi}_k$ on any category k varies within $[10^{-1}, 10^1, \ldots, 10^{13}]$. To calculate the weight vector $\boldsymbol{w}$, we learn it by applying the linear regression to the corresponding statistical result from all related data contributors, and further normalize the vector so as to get unit L2 norm. For the weighted sum query by the perturbed matrix, we evaluate the accuracy performance and the inference probability of *PEATSE* and existing work [27] for sensitive and insensitive categories, respectively.

Specifically, the metrics for accuracy performance include the data consumer's accuracy and the probability of the query answer belonging to the acceptable range $[\zeta - \boldsymbol{\delta}_k, \zeta + \boldsymbol{\delta}_k]$ (also called acceptable probability). Moreover, the accuracy is defined as $\alpha = 1 - \frac{|\hat{\zeta} - \zeta|}{|\hat{\zeta} + \zeta|}$, where ζ and $\hat{\zeta}$ denote the true value by the initial matrix and the weighted sum query answer according to the perturbed matrix, respectively. In addition, each data point is the average report after running 200 times.

4.2 Evaluation Results

4.2.1 Evaluation of Accuracy. Fig.2 and Fig.3 demonstrate the data consumer's accuracy both goes up when the privacy preference $\boldsymbol{\phi}_k$ increases for sensitive and insensitive categories, respectively, which proves usefulness of *PEATSE*. The reason lies in the fact that the higher privacy preference leads to a smaller variance $2\lambda^2$, and probably generates smaller Laplace noise. Thus, the query answer from the perturbed matrix would not deviate from the true value largely.

In Fig.3, for both Zhang *et al.*'s work [27] and *PEATSE*, we can observe that the accuracy results are close to each other under the same dimensionality $d = 3$ or $d = 15$ especially for the privacy preference larger than 10^9. Obviously, the reason is also two approximate query answers due to smaller noise. *PEATSE* is inferior to Zhang *et al.*'s work [27] slightly for smaller privacy preference, because they sacrifice the data contributor's privacy by enlarging the privacy budget to $\boldsymbol{\phi}_k * \omega(V_i, V_j)$ in spite of the higher data accuracy.

4.2.2 Evaluation of Acceptable Probability. From Fig.4 and Fig.5, it can be seen that the acceptable probability goes up with the increase of the error bound $\boldsymbol{\delta}_k$, for both sensitive and insensitive categories, which guarantees each data consumer can obtain the acceptable query answer with a higher probability, and further reflects feasibility of *PEATSE*. It is understood that the higher error bound leads to a broader acceptable range for the data consumer, and thus the weighted sum query answer belongs to this range with a higher probability.

Specifically, *PEATSE* has a higher acceptable probability than existing work [27] under any fixed error bound $\boldsymbol{\delta}_k$ when $\boldsymbol{\delta}_k$ ranges within $[1, 2.4]$. This is because existing work performs unsteadily when the error bound is smaller, and is inadequate to solve the trading problem with the data consumer's accuracy requirement. In addition, it appears a higher acceptable probability for the higher dimensionality $d = 15$ than the lower dimensionality $d = 3$ for any fixed $\boldsymbol{\delta}_k$, which shows added noise would not increase for the higher dimensionality, and also verifies validity of *PEATSE*.

4.2.3 Evaluation of Total Compensation. In Fig.6, we observe that the data broker pays less privacy compensation for chosen data contributors when the error bound $\boldsymbol{\delta}_k$ increases from 1 to 2.4. It is obvious that the higher error bound means the smaller privacy loss according to Algorithm 1, and thus the compensation becomes less by the payment scheme.

4.2.4 Evaluation of Inference Probability. Fig.7 shows both *PEATSE* and existing work [27] achieve privacy protection when the attacker only has limited prior information for the victim, where the inference probability represents the probability of an attacker with given information inferring the victim's browsing records from the perturbed matrix. Without loss of generality, the attacker's prior information is modeled as known fractional data elements, where the number d varies within the range $[1, 25]$. In addition, for any victim, we set the attacker's known data vector $\hat{V}_i$ as the vector of known d elements and zero for other unknown features.

Next, we compare the distance between $\hat{V}_i$ and each data vector from the initial matrix and perturbed matrix, respectively. It can be observed that the inference probability is significantly improved at dimensionality $d = 13$ for the initial matrix. This is because the fewer data contributors' data vectors are close to the known vector $\hat{V}_i$, and the victim's real identity would be exposed with the higher probability with the increase of dimensionality. However, *PEATSE* shows the lower inference probability for the perturbed matrix, and indeed protects the data contributor's privacy even when dimensionality increases to $d = 22$. In addition, we can see *PEATSE* outperforms slightly than Zhang *et al.*'s work [27] about the inference probability from Fig.7.

According to the above evaluation results, *PEATSE* guarantees the data consumer's accuracy requirement and data contributors' privacy preferences simultaneously, and well balances the utility and user privacy.

5 RELATED WORK

5.1 Query-based Trading

A growing number of related literatures [15][6][13][16] have investigated query-based trading for data markets like Acxiom [4] in recent years. Koutris *et al.* [13] first propose query-based pricing especially in consideration of the buyer's possible arbitrage behavior. The buyer possibly infers more accurate query answer by asking multiple cheaper queries with the less query cost. However, they focus on general-purpose data rather than private data. Conversely, follow-up work by Li *et al.* [15] and Niu *et al.* [20] also design arbitrage-free pricing function, but further consider privacy loss from data contributors when releasing perturbed common aggregated statistical results about the population, as well

as the corresponding privacy compensation mechanism design. Specifically, only the query answer by added Laplace noise is returned to the buyer.

However, different from the above work, we focus on the trading of the entire dataset rather than the single query answer in terms of most analysts' preferences for the noisy dataset, also regarded as the non-interactive model. Therefore, our work seriously perturbs each data element in the returned dataset in case of the possible user-linkage attack [26]. In addition, it becomes more difficult to supply the utility of the released dataset simultaneously. Specifically, Zhang *et al.* [27] only design the data perturbation mechanism for the text-based Twitter dataset, but neglect the analysis for the buyer's utility by the purchased perturbed dataset. On the contrary, our work achieves accuracy guarantee by the rigorous proof when analyzing the performance of the returned dataset for the data consumer.

5.2 Incentive Issues for Trading

Other previous work [11][24][9] mainly aim at incentivizing data contributors to disclose their real privacy valuations in the context of mechanism design. Under the assumption of the unknown privacy valuation, each data contributor probably misreports a higher privacy valuation for the higher benefit. Ghosh *et al.* [11] regard privacy as traded commodity, and design the privacy trading mechanism for the counting query by running truthful auction. Wang *et al.* [24] propose an incentive mechanism to make data contributors control their own data privacy by reporting a noisy version and the truthful valuation in terms of the untrusted data collector. Moreover, Nissim *et al.* [19] further consider the data contributor's price for privacy as private information because the higher price probably reflects his more sensitive information.

Unfortunately, none of the above work has taken data contributors' diverse privacy preferences on various kinds of web browsing histories into consideration, and further considered the trading of the entire dataset and the utility of the returned dataset.

ACKNOWLEDGMENTS

This research is supported in part by NSFC (No. 61772341, 61472254) and STSCM (No. 18511103002). This work is also supported by the Program for Changjiang Young Scholars in University of China, the Program for China Top Young Talents, the Program for Shanghai Top Young Talents, Shanghai Engineering Research Center of Digital Education Equipment, and SJTU Global Strategic Partnership Fund (2019 SJTU-HKUST).

6 CONCLUSIONS

In this paper, we have proposed a privacy-preserving trading framework *PEATSE* for entire web browsing histories in consideration of both data contributors' diverse privacy preferences and the data consumer's utility. The data consumer can purchase the perturbed user-by-features matrix with guarantee of the desired accuracy requirement. Besides, data contributors can be compensated for privacy loss due to the delivery of sensitive and insensitive browsing records, and they have to report private costs truthfully for maximizing their utility. Through real-data based experiments, the evaluation and analysis results demonstrate *PEATSE* well balances user privacy and the data consumer's utility, and further achieves desirable economic properties of truthfulness, individual rationality and budget balance.

REFERENCES

[1] 2011. Forecast of big data market size, based on revenue, from 2011 to 2026. "http://www.statista.com/".
[2] 2012. Infochimps. "https://www.infochimps.com/marketplace.".
[3] 2014. Data Brokers: A Call For Transparency and Accountability: A Report of the Federal Trade Commission. "https://www.ftc.gov/reports/".
[4] 2017. Behavior Targeting. "https://business.twitter.com/en/targeting.html.".
[5] Ghazaleh Beigi, Ruocheng Guo, Alexander Nou, Yanchao Zhang, and Huan Liu. 2018. Protecting User Privacy: An Approach for Untraceable Web Browsing History and Unambiguous User Profiles. *CoRR* (2018).
[6] Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *SIGACT-SIGMOD-SIGART 2003, June 9-12, 2003, San Diego, CA, USA*. 202–210.
[7] Cynthia Dwork. 2011. *Differential Privacy*. 338–340 pages.
[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Third Theory of Cryptography Conference, New York, NY, USA, March 4-7*. 265–284.
[9] Lisa Fleischer and Yu-Han Lyu. 2012. Approximately optimal auctions for selling privacy when costs are correlated with data. In *EC 2012, Valencia, Spain, June 4-8*. 568–585.
[10] Robert S. Garfinkel, Ram D. Gopal, Manuel A. Nunez, and Daniel O. Rice. 2006. Secure electronic markets for private information. *IEEE Trans. Systems, Man, and Cybernetics, Part A* 36, 3 (2006), 461–471.
[11] Arpita Ghosh and Aaron Roth. 2015. Selling privacy at auction. *Games and Economic Behavior* 91 (2015), 334–346.
[12] Zach Jorgensen, Ting Yu, and Graham Cormode. 2015. Conservative or liberal? Personalized differential privacy. In *ICDE 2015, Seoul, South Korea, April 13-17*. 1023–1034.
[13] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2013. Toward practical query pricing with QueryMarket. In *SIGMOD 2013, New York, NY, USA, June 22-27*. 613–624.
[14] Kenneth C. Laudon. 1996. Markets and Privacy. *Commun. ACM* 9 (1996), 92–104.
[15] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. 2017. A theory of pricing private data. *Commun. ACM* 60, 12 (2017), 79–86.
[16] Chao Li and Gerome Miklau. 2012. Pricing Aggregate Queries in a Data Marketplace. In *Proceedings of the 15th International Workshop on the Web and Databases 2012, Scottsdale, AZ, USA, May 20*. 19–24.
[17] Rachana Nget, Yang Cao, and Masatoshi Yoshikawa. 2017. How to Balance Privacy and Money through Pricing Mechanism in Personal Data Market. In *SIGIR 2017 Workshop, Tokyo, Japan, August 11*.
[18] Noam Nisan and Roughgarden. 2007. *Algorithmic game theory*. Cambridge University Press Cambridge.
[19] Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. 2012. Privacy-aware mechanism design. In *EC 2012, Valencia, Spain, June 4-8, 2012*. 774–789.
[20] Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Shaojie Tang, Xiaofeng Gao, and Guihai Chen. 2018. Unlocking the Value of Privacy: Trading Aggregate Statistics over Private Correlated Data. In *KDD 2018, London, UK, August 19-23*. 2031–2040.
[21] Do Viet Phuong and Tu Minh Phuong. 2013. Gender Prediction Using Browsing History. In *KSE 2013, Volume 1, Hanoi, Vietnam, 17-19 October*. 271–283.
[22] Klaus Schwab, Alan Marcus, JO Oyola, William Hoffman, and M Luzi. 2011. Personal data: The emergence of a new asset class. In *An Initiative of the World Economic Forum*.
[23] Eran Toch, Yang Wang, and Lorrie Faith Cranor. 2012. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction* 1-2 (2012), 203–220.
[24] Weina Wang, Lei Ying, and Junshan Zhang. 2016. The Value of Privacy: Strategic Data Subjects, Incentive Mechanisms and Fundamental Limits. In *SIGMETRICS 2016, Antibes Juan-Les-Pins, France, June 14-18*. 249–260.
[25] Bin Yang, Issei Sato, and Hiroshi Nakagawa. 2015. Bayesian Differential Privacy on Correlated Data. In *SIGMOD 2015, Melbourne, Victoria, Australia, May 31 - June 4*. 747–762.
[26] Xianqi Yu, Yuqing Sun, Elisa Bertino, and Xin Li. 2018. Modeling user intrinsic characteristic on social media for identity linkage. *ACM Transactions on Social Computing* 3 (2018), 11.
[27] Jinxue Zhang, Jingchao Sun, Rui Zhang, Yanchao Zhang, and Xia Hu. 2018. Privacy-Preserving Social Media Data Outsourcing. In *INFOCOM 2018, Honolulu, HI, USA, April 16-19*. 1106–1114.
[28] Zhenzhe Zheng and Yanqing Peng. 2017. Trading Data in the Crowd: Profit-Driven Data Acquisition for Mobile Crowdsensing. *IEEE Journal on Selected Areas in Communications* 35 (2017), 486–501.