Published in final edited form as:

ACM Trans Manag Inf Syst. 2019 July ; 10(2): . doi:10.1145/3329717.

Differential Privacy for the Vast Majority

HASAN B. KARTAL,

Department of Management Information Systems, College of Business and Management,
University of Illinois at Springfield, One University Plaza, UHB 4030, Springfield, IL 62703

XIAOPING LIU,

D'Amore-McKim School of Business, Northeastern University, 360 Huntington Ave, Boston, MA
02115

XIAO-BAI LI

Department of Operations and Information Systems, Manning School of Business, University of
Massachusetts Lowell, Pulichino Tong Business Center, 448, Lowell, MA 01854

Abstract

Differential privacy has become one of the widely used mechanisms for protecting sensitive information in databases and information systems. Although differential privacy provides a clear measure of privacy guarantee, it implicitly assumes that each individual corresponds to a single record in the result of a database query. This assumption may not hold in many database query applications. When an individual has multiple records, strict implementation of differential privacy may cause significant information loss. In this study, we extend the differential privacy principle to situations where multiple records in a database are associated with the same individual. We propose a new privacy principle that integrates differential privacy with the Pareto principle in analyzing privacy risk and data utility. When applied to the situations with multiple records per person, the proposed approach can significantly reduce the information loss in the released query results with a relatively small relaxation in the differential privacy guarantee. The effectiveness of the proposed approach is evaluated using three real-world databases.

CCS Concepts:

Information systems → Database query processing; Security and privacy → Privacy protections;
Social and professional topics → Privacy policies

Additional Key Words and Phrases:

Data privacy; database query; Pareto principle; noise perturbation

1 INTRODUCTION

Large volumes of data are collected and stored on a daily basis by organizations, and data collected about individuals and transactions are growing exponentially (Tuarob et al. 2018).

While the availability of this data can provide rich benefits to organizations and society at large, there is a growing concern about data privacy by consumers. When shared data include personal information, it is required that appropriate measures be taken to protect the privacy of the individuals. Traditional approaches commonly used for privacy protection, such as k -anonymity (Sweeney 2002; Fu et al. 2010) and noise-based perturbation (Aggarwal and Yu 2008), typically depend on some assumptions about the privacy adversary's auxiliary information regarding individual subjects. When the assumptions are violated, these approaches may not work well.

To understand the issue, consider a patient dataset that includes the patients' demographic attributes such as age, gender, and zip code, as well as some medical- and healthcare-related attributes. In order to make the data available for external users, the data owner organization who adopts a traditional privacy approach needs to first make assumptions about what background or auxiliary information about the patients an adversary may have known. For example, the data owner may assume that the adversary knows the age, gender, and zip code of some patients. Studies (Sweeney 2002; Aggarwal and Yu 2008) have found that it is not difficult to uniquely identify a person using age, gender, and zip code information. Therefore, the data owner would alter the values of these three attributes when making the data available for access (by generalizing the attribute values if using k -anonymity, or adding noise to the attribute values if using a noise perturbation method). If the adversary actually does not know, say, the age of the target patients, then changing the age values causes unnecessary information loss. On the other hand, if the adversary knows some information that is not considered as known by the data owner, e.g., a specific health insurance plan, then the patients with that information in the data will be subject to higher disclosure risks.

To overcome this limitation, Dwork (2006, 2011) introduced the notion of differential privacy, which assesses privacy disclosure risk without any assumptions about the adversary's background knowledge. Differential privacy principle holds that the privacy risk of an individual should not substantially change as a result of the individual's participation in a database. Differential privacy mechanism is then built to ensure that the released information about a database is essentially the same whether or not an individual's data is included in the database. In other words, there is virtually no additional privacy disclosure risk if the individual opts into the database. Consider again the patient data example. Suppose a database query attempts to find the number of patients having specified values of age, gender, zip code, and health insurance plan. Differential privacy will answer the query by adding a certain amount of noise to the true count such that perturbed numbers will be about the same whether or not a patient with the specified values of age, gender, zip code, and health insurance plan is included in the database. This is achieved without any assumption regarding the adversary's background knowledge about the patients' age, gender, zip code, and health insurance plan.

Because differential privacy is defined independent of any auxiliary information assumption, it provides the most rigorous privacy guarantee among existing approaches. There have been many recent implementations of differential privacy for real-world applications. Google has adopted differential privacy to collect usage data from its Chrome users with privacy protection (Erlingsson et al. 2014). Apple has started using differential privacy in its iOS10

to find the usage patterns of a large number of users without compromising user privacy (Apple Inc. 2016). In July 2017, Uber released an open-source tool for data sharing based on differential privacy principle (Uber Inc. 2017). Furthermore, in September 2017, the U.S. Census Bureau announced that differential privacy mechanism will be applied to prototype end-to-end census test products, and the mechanism will be fully implemented for the 2020 census of U.S. population (U.S. Census Bureau 2017).

Although differential privacy provides a clear measure of privacy guarantee, it treats data utility as a secondary objective, considered only after privacy criteria are rigorously satisfied. Consequently, existing differential privacy mechanisms often result in significant information loss, which limits its application domains. Thus, although there is an increasing number of differential privacy-based applications, researchers have also found many cases where it is practically impossible to achieve a reasonable utility if differential privacy principle must be strictly satisfied (Machanavajjhala et al. 2008; Dankar and El Emam 2012). This is indeed the case when an individual has multiple records in a database. Differential privacy implicitly assumes that each individual corresponds to a single record (Dwork 2006, 2011). However, in many real-world databases, such as patient diagnosis records, phone call records, credit card transactions, and web page clickstreams, an individual often corresponds to multiple records (El Emam et al. 2013). In this situation, the distribution of individuals' occurrences can be highly unbalanced, and strict implementation of differential privacy can cause substantial information loss. To our knowledge, this problem has not been studied in the data privacy literature.

Now, assume that the patient data example contains each visit of a patient as a record and we consider answering a count query on the database. Differential privacy ensures that, by perturbing the count output with a certain amount of noise, the presence or absence of a patient in the database will not significantly change the output. The amount of noise to be added to the output, however, depends on the context of count query. If the query is to find the *number of patients* having a certain disease and located in a certain zip code area, then the amount of added noise is minimal, and the perturbed output will be very close to the true result. However, if the query is to find the total *number of visits* of the patients having that disease and located in that zip code area, then adding or removing a patient may change the outcome of the count query by more than one count as some patients can have multiple visits. Consequently, the amount of added noise can be large (in both absolute and relative terms). Specifically, if there are a few patients with a lot of visits, the perturbed output can depart significantly from the true result, because the noise added to the output by differential privacy depends on the largest number of visits from a patient.

We have used count query to illustrate the advantages and limitations of using differential privacy. Count query is one of the very basic data analysis functions. It is widely used in descriptive analytics not only for providing basic counting statistics but also for generating various summary representations such as histograms, distributions, column and pie charts, and data visualization plots. Count function is also used in predictive analytics, e.g., for finding large itemset count for mining association rules. Furthermore, it is frequently used in unstructured data analysis, e.g., for online keyword search count and for constructing the term-frequency matrices for text mining (Berndt et al. 2015). In these applications, the count

values are calculated in terms of specified criteria. If the reported count value is small, the individuals whose attribute values match the specified criteria are subject to high disclosure risk. In this article, we focus on disclosure risk and data utility issues when using differential privacy for count function. However, the basic idea of our research can be extended to many other data analysis functions.

This work proposes a new privacy principle based on the ideas of differential privacy and the Pareto principle. The proposed privacy principle considers the tradeoff between privacy and utility and helps the data owner organization to choose an appropriate amount of noise for perturbing count query output with some easy to interpret parameters. The proposed principle, when applied to a database with multiple records per person, significantly improves the accuracy of the query results, and also guarantees privacy at a level defined by a parameter controlled by the data owner. The effectiveness of the proposed approach is evaluated using three real-world databases.

2 BACKGROUND

In this section, we provide a preliminary overview of differential privacy mechanism and discuss information loss issue associated with differential privacy when dealing with highly unbalanced occurrence distribution.

2.1 Differential Privacy

Differential privacy is defined in terms of the impact of adding (or removing) one individual to (or from) a database on the result of query output or data analysis. Let D_1 and D_2 be the dataset before and after adding or removing the individual, respectively. The definitions in this subsection are based on Dwork (2006).

Definition 1 (Dwork 2006). Given any two datasets D_1 and D_2 that differ in only one record, a randomized perturbation mechanism M provides ϵ -differential privacy (ϵ -DP) if for any set of possible outputs S of M (i.e., $S \subseteq \text{Range}(M)$)

$$\Pr[M(D_1) \in S] \leq e^\epsilon \times \Pr[M(D_2) \in S] \quad (1)$$

The parameter ϵ represents disclosure risk, which is usually controlled to be small so that e^ϵ is close to 1. As such, differential privacy guarantees, in a probabilistic sense, that the outputs will be essentially the same with or without any specific individual's participation. This property has a very appealing implication. For example, if the dataset were to be used by a healthcare provider to analyze the demographics of its patient population, then the presence or absence of a patient's record in the dataset will not significantly change the results of the analysis (e.g., count query for demographics attributes). In this sense, the participating patient's demographic information is well hidden. Differential privacy mechanism uses a measure called sensitivity defined below.

Definition 2 (Dwork 2006). For a function f over dataset D with numeric output, the sensitivity of f is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \quad (2)$$

for all D_1 and D_2 differing in at most one record.

In other words, the sensitivity is the maximum change in the value of f when any single record of D changes (in a single record per person scenario). It has been shown (e.g., in Dwork (2011)) that *for a numeric function f , a perturbation mechanism that adds noise with a Laplace (f/ϵ) distribution to the output satisfies ϵ -DP*. The Laplace (σ) distribution has a density function of

$$p(x / \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right).$$

When f represents a count query, sensitivity $f = 1$ since the count can differ at most by one due to the addition or removal of one record (again, in a single record per person scenario). Then, for the count query f , ϵ -DP is achieved by the following randomized noise addition mechanism:

$$M(D) = f(D) + \text{Laplace}(1 / \epsilon) \quad (3)$$

When an individual has multiple records in a database, the presence or absence of an individual can change the value of a count function by more than one if the count result is calculated based on the number of occurrences of the individuals. One approach in this situation is to evaluate the sensitivity of the count function based on the number of records an individual has. Suppose the maximum number of records of any individual in a database is k . Then the sensitivity of a count function becomes $f = k$, which increases the amount of added noise by k times. This would have a significantly negative impact on the accuracy of the count query output.

2.2 Information Loss When Using Differential Privacy for Highly Unbalanced Occurrence Distribution

In differential privacy, the added noise is proportional to the sensitivity of the query function, which is determined by the individual having the largest impact on the output of the function. However, often most of the remaining individuals in the database have a much smaller impact on the output. So, a single individual or a small proportion of the individuals in a database can cause a significant increase in noise addition. For count query involving multiple occurrences of an individual (e.g., number of patient visits), an individual with a large number of occurrences requires a large added noise, causing a significant amount of information loss in the output of a count query.

To illustrate the problem, consider the examples in Figure 1, which shows three different scenarios for 10 individuals with different numbers of occurrences. The horizontal axis represents each individual, and the vertical axis or bins represents the number of records

(occurrences) for the individuals. Suppose a count query is issued: “what is the total number of all records (occurrences) in the database?” The sensitivity of the count query is equal to the largest number of occurrences, which is 10 for all three scenarios. Hence, a differential privacy mechanism would add the same amount of noise to the query output for each of the three scenarios, even though the total number of records in each scenario is quite different.

Using Equations (2) and (3), Table 1 shows the query output results with $\epsilon = 2$ where the output range of the answers to the query are calculated based on a 98% confidence interval and $f/\epsilon = 10/2 = 5$. It is evident from Table 1 that the amount of noise added relative to the actual output is much higher for the data in Figure 1(a) than those in the other two scenarios. This is because, in Figure 1(a), the number of records of the first individual is significantly (10 times) larger than those of all the remaining individuals. On the other hand, the amount of added noise is minimal in Figure 1(c) because all individuals have the same number of records. Therefore, the standard differential privacy mechanism should work well for the data in Figure 1(c) but not that in Figure 1(a).

To address the problem illustrated in Figure 1(a), our idea is to calibrate noise addition based on the Pareto principle so that with a relatively small relaxation in ϵ -DP we can significantly reduce the information loss in output.

3 PARETO PRINCIPLE FOR DIFFERENTIAL PRIVACY

3.1 Pareto Principle and Distribution

In many business and socioeconomic data, one can observe that a large proportion of the effects are owing to a small proportion of the causes. For example, 20% of the population holds 80% of society’s wealth. Similarly, 20% of the customer accounts for 80% of the total revenue. This is generally known as the “80/20 rule.” The Pareto principle generalizes this rule into “ $(1 - p)/p$ law” for some $p \in (0, 0.5)$. The Pareto rule is observed quite often and pertains to many other statistical distributions. In order to describe our proposed approach, which is related to Pareto distribution, we provide definitions for Pareto distribution and generalized Pareto distribution below.

Definition 3 (Pareto Distribution). For a random variable X , the *Pareto distribution* has the following cumulative distribution function $F(x)$, and the probability density function $f(x)$:

$$F(x) = \Pr(X \leq x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^\alpha, & x \geq x_m \\ 0, & x < x_m \end{cases} \quad (4)$$

$$f(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m \\ 0, & x < x_m \end{cases} \quad (5)$$

where the shape parameter $\alpha > 0$ is called the *Pareto index*, and the location parameter x_m is the minimum possible value of X .

Definition 4 (Generalized Pareto Principle (Hardy 2010)). For any real $r \geq 0$, there is a unique $p \in (0, 1)$ such that $\Pr(X > x) = p^r$, and for that value we have

$$\frac{\int_{X > x} X dv}{\int_{\Omega} X dv} = (1 - p)^r \quad (6)$$

where “ $(X > x)$ ” means $\{\omega \in \Omega: X(\omega) > x\}$. In this case, X is said to follow a *generalized Pareto distribution*.

It has been shown (Hardy 2010) that, for $1 < \alpha < \infty$, the Pareto distribution as characterized by Equation (4) or Equation (5) satisfies the generalized Pareto principle. In this case, α in the Pareto distribution and p in the generalized Pareto distribution has the following relationship:

$$\alpha = \log_p / (1 - p) p > 1. \quad (7)$$

The generalized Pareto principle can be stated as “ $(1 - p)^r \times 100\%$ of the effects come from $p^r \times 100\%$ of the causes.”

3.2 The Proposed (ϵ, p') -Differential Privacy Principle

As described in Section 2, when using differential privacy for returning the results of count queries on a database where individuals have multiple records, the added noise can be unreasonably large due to a few high-frequency individuals, causing huge information loss that renders the results practically useless. To address this problem, our idea is to apply Pareto principle to separate a large number of low-frequency individuals from a small number of high-frequency ones so that information loss can be significantly reduced when the sensitivity for differential privacy is calculated based on the group of low-frequency individuals. To characterize our idea rigorously, we formally state our proposed privacy principle, called (ϵ, p') -differential privacy, in Definition 5 below.

Definition 5 ((ϵ, p') -Differential Privacy). Given a dataset D , a randomized perturbation mechanism M satisfies (ϵ, p') -differential privacy $((\epsilon, p')$ -DP) if we can divide D into two non-overlapping sets, D^A and D^B , such that

$$|D^B| \leq p^r |D|, \text{ or equivalently, } |D^A| \geq (1 - p^r) |D|, \quad (8)$$

where $|D|$, $|D^A|$, and $|D^B|$ are the number of individuals in dataset D , D^A , and D^B , respectively; and if for all $D_1^A \subseteq D^A$ and $D_2^A \subseteq D^A$ that differ in at most one individual and for any possible output space S generated from D^A ,

$$\Pr[M(D_1^A) \in S] \leq e^\epsilon \times \Pr[M(D_2^A) \in S]. \quad (9)$$

The (ϵ, p^r) -DP ensures that perturbation mechanism M satisfies ϵ -DP for at least $1 - p^r$ proportion of individuals. Parameter r is a user specified parameter (often we can set $r = 1$); and p is generally small ($P \ll 0.5$). So, (ϵ, p^r) -DP ensures that M satisfies ϵ -DP for a vast majority of the individuals. For a small proportion (p^r), M cannot satisfy ϵ -DP. Instead, it provides a lower level of privacy protection by adding a noise determined by the sensitivity of the majority set (D^A). This is the tradeoff in order to avoid a significant loss in the accuracy of query output.

For the count function f , let Δf and $\Delta f_{p,r}$ be the sensitivity of f for ϵ -DP and (ϵ, p^r) -DP, respectively. It follows from Definition 5 that $\Delta f > \Delta f_{p,r}$. With noise generated from a Laplace distribution, the variance of the noise added by ϵ -DP and (ϵ, p^r) -DP is $2(\Delta f / \epsilon)^2$ and $2(\Delta f_{p,r} / \epsilon)^2$, respectively. Thus, the proportion of variance reduced by (ϵ, p^r) -DP relative to ϵ -DP is

$$\frac{2(\Delta f / \epsilon)^2 - 2(\Delta f_{p,r} / \epsilon)^2}{2(\Delta f / \epsilon)^2} = 1 - \left(\frac{\Delta f_{p,r}}{\Delta f} \right)^2 \quad (10)$$

So, the proportion of variance reduced does not depend on ϵ and it is a quadratic function of the change in sensitivity that varies with r .

3.3 Achieving (ϵ, p^r) -Differential Privacy

We explain how to achieve (ϵ, p^r) -differential privacy based on the Pareto principle using a dataset with count values distributed similarly to the example in Figure 1(b). The occurrence distribution of the individuals is shown in Figure 2, where each shaded horizontal bar represents the additional loss or gain in the accuracy of the total occurrence count if the corresponding individual has the largest number of occurrences and is added to or removed from the dataset. For example, the top horizontal bar represents the accuracy gain if individual #1 is removed from the dataset. The second top horizontal bar represents the accuracy gain if individual #2 is removed after removing #1. Two individuals are said to be *differentially adjacent* if they are adjacent in the sorted list and they have a different number of occurrences. For example, individuals #1 and #2 are differentially adjacent, as well as #2 and #3. Individuals #3 and #4 are not differentially adjacent because they have the same number of occurrences.

Let f_i be the number of occurrences of the i th individual and n_i be the number of individuals who have a smaller number of occurrences than what the i th individual has. Let t_i be the number of individuals who have the same number of occurrences as the i th individual (including the i th individual), which is greater than one when the i th individual is non-differentially adjacent (tied) to some other individuals. The change in count accuracy due to the addition or removal of the i th individual, when it is on top of the sorted list, is

$$q_i = n_i(f_i - f_{i+1}) / t_i \quad (11)$$

In Figure 2, this quantity is represented by the area of the shaded rectangle adjacent to the right of the respective bin. For example,

$$q_1 = n_1(f_1 - f_2) / t_1 = 9(46 - 32) / 1 = 126.$$

After removing individual #1, we have

$$q_2 = 8(32 - 20) / 1 = 96.$$

Individuals #3 and #4 are not differentially adjacent. So,

$$q_3 = q_4 = 6(20 - 16) / 2 = 24 / 2 = 12.$$

Figure 2 shows that if we apply a (ϵ, p^f) -DP mechanism by separating the first two individuals from the remaining eight individuals, then 80% of the accuracy loss caused by ϵ -DP can be recovered.

For a given dataset of n individuals with $\mathbf{q} = (q_1, q_2, \dots, q_n)$, we can estimate the Pareto index α from $\mathbf{q}$ using a maximum likelihood method. Let $q_m = \min q_i$; the maximum likelihood estimator for α is (Arnold 2015)

$$\hat{\alpha} = n / \sum_i (\ln q_i - \ln q_m) \quad (12)$$

After α is estimated, the value of parameter p can be computed based on Equation (7). Then, the noise perturbation can be implemented on the query output to satisfy (ϵ, p^f) -DP.

The (ϵ, p^f) -DP ensures that at least the $1 - p^f$ proportion of individuals are protected with ϵ -DP (while the other p^f proportion of individuals are protected at a lower level). On the other hand, based on Definition 4 the proportion of loss in count accuracy recovered by (ϵ, p^f) -DP is $(1 - p)^f$. While the value of parameter p is computed from data, parameter r is determined by the data owner to specify different tradeoffs between the proportion of individuals ensured by ϵ -DP and the proportion of recovered loss in the accuracy of query result. For the illustrative example in Figure 2, the computed value of p is approximately 0.2. Table 2 shows the tradeoff between ϵ -DP protection and accuracy recovery with several different r values. It is clear that a small relaxation in ϵ -DP can lead to a significant gain in output accuracy.

4 EXPERIMENTAL EVALUATION

We evaluate our proposed approach using three real-world databases. The first is a patient database provided by the INFORMS Data Mining Section (2008), which includes 117,307 patient visits from 29,531 patients. The database consists of four related datasets, two of which are closely related to multiple-records problems: Patient Demographics and Patient Conditions. The Patient Demographics dataset includes attributes such as patient ID, year of birth, gender, race, years of education, marital status, and income level. The second is a movie rating database collected by Harper and Konstan (2016), which contains 100,000 user ratings submitted by 943 users. The database consists of three related datasets: User, Rating, and Information datasets. The User dataset provides demographic information for the users such as user ID, age, gender, occupation, and zip code. The Rating dataset contains the user ratings (1–5 scale) for 1,682 movies with timestamp. The Information dataset contains movie information such as movie ID, name, and genre. The third is a financial database owned by a bank (PKDD 1999), containing 273,508 transaction records from 3,674 clients. The database consists of eight related datasets such as Client, Account, Demographics, and Transactions.

Figures 3, 4, and 5 display the overall distributions for the individuals in the three databases, where the horizontal axis represents the individuals in each database and the vertical axis represents the number of patient visits, the number of movie ratings, and the number of transactions, respectively. Clearly, in each database, a small proportion of the individuals account for a large proportion of occurrences. For the overall occurrence distribution of the individuals, the estimated value of parameter p is 0.1037 for the Patient dataset, 0.1796 for the Movie dataset, and 0.2192 for the Bank dataset.

To perform the experiments, we needed to determine the values of parameter ϵ first. Parameter ϵ represents the difference between the probabilities of getting the same output from two datasets that differ in one individual. This difference is stated in terms of the log of probability, as seen in Equation (1), which is not easy to interpret directly. Essentially, the selection of ϵ value depends on the actual data and application context (Lee and Clifton 2011). In general, Dwork (2011) suggests that the ϵ value be between $\ln(1.01) \approx 0.01$ and $\ln(3) \approx 1.10$. A smaller ϵ value implies a stronger privacy protection. In practice, Apple has used ϵ values of between 2 and 8 in various contexts (Apple 2017). Based on the values suggested in the general guidelines and used in practice, we tested some ϵ values to identify the values leading to both fairly small disclosure risk and reasonable data utility. Consequently, we selected $\epsilon = \ln(1.5) \approx 0.41$ for the Patient data and $\epsilon = \ln(3) \approx 1.10$ for the Movie and Bank data. The ϵ values for the Movie and Bank data are somewhat large because their sensitivity values f are relatively large. The added noise will be too large for the Movie and Bank data if the selected ϵ values are too small.

We should point out that the choice of ϵ value will not really affect the results of our experimental evaluation, which compares our proposed (ϵ, p') -DP approach with the standard ϵ -DP approach based on the same ϵ value. The focus of our experimental evaluation is on the tradeoff between the proportion of ϵ -DP ensured individuals under the

proposed approach and related changes in sensitivity and variance of the noise. We have explained with Equation (10) that the measures related to the tradeoff are independent of ϵ .

Our evaluation study focused on count query. For each database, we run three count queries for each database, resulting in a total of nine queries in the experiment. We used Equations (12) and (7) to compute p values in each query and reported them along with the selected r values for (ϵ, p') -DP. We then compare the proposed (ϵ, p') -DP approach with the standard ϵ -DP mechanism in terms of sensitivity of the query, the 95% confidence interval of the noise added to the output, and the variance of added noise. We also examine, for different r values in (ϵ, p') -DP, the tradeoff between the proportion of individuals with ϵ -DP ensured and the changes in data utility, including those in sensitivity, output range, and variance of the noise.

(A) Queries on the Patient Database.

The users of the patient database may be interested in the number of patient visits made by older people and/or by low-income people. Therefore, we have experimented with three count queries, first conditioned on age, the second conditioned on income, and the third on both age and income. The results are reported in Tables 3, 4, and 5.

Query A1: Number of patient visits where patient is 65 years old or older.

Query A2: Number of patient visits where patient income is less than \$15,000.

Query A3: Number of patient visits where patient is 65 years old or older AND income is less than \$15,000.

As we can see from Tables 3–5, with ϵ -DP the added noise for the three queries can be several times larger than the original values, resulting in output values that are likely to deviate significantly from the true values. On the other hand, the proposed (ϵ, p') -DP approach yields much more useful results. For a small reduction in the proportion of the individuals ensured with ϵ -DP, there is a large return in the reduced variance of noise by (ϵ, p') -DP. As r increases, the proportion of the individuals ensured with ϵ -DP increases, but the loss in output accuracy, as characterized by the range of confidence interval, increases at a faster rate. It appears that $r = 1.5$ or so is a reasonable choice, which results in more than 93.6% of the patients ensured with ϵ -DP, together with much more accurate query output.

It is important to note that the perturbed values are generated with Laplace noise, which follows a mountain-shaped distribution. Thus, although the confidence intervals look wide in range, most values of the added noise are much smaller than the boundary values, and the perturbed count values are mostly centered on the original values.

(B) Queries on the Movie Database.

The users of the movie rating database may be interested in the number of movie ratings of a certain genre and/or those with a high score. Therefore, we have tested with three count queries: the first conditioned on the drama genre, the second conditioned on the highest score of 5, and the third on both the drama genre and the score of 5. The results are reported in Tables 6, 7, and 8.

Query B1: Number of user ratings where movie genre is drama.

Query B2: Number of user ratings where rating score is 5.

Query B3: Number of user ratings where movie genre is drama AND rating score is 5

As we can see from Tables 6–8, with ϵ -DP the 95% confidence intervals covering the added noise for the three queries are approximately 3 to 15 times larger than the original values. With this result, it is really questionable if ϵ -DP is really useful. On the other hand, (ϵ, p') -DP results in much smaller added noise. It appears that $r = 1.5$ or so is a reasonable choice, which results in more than 93.7% of the patients ensured with ϵ -DP, together with much more accurate query output. In general, it is more difficult to apply ϵ -DP or (ϵ, p') -DP to the movie rating data than to the patient data.

(C) Queries on the Bank Database.

The bank database users may be interested in the number of transactions made on weekends and/or those with a relatively large amount. Therefore, we have tested with three count queries: the first conditioned on weekend transactions, the second conditioned on the amount larger than \$10,000, and the third on transactions on weekends with an amount larger than \$10,000. The results are reported in Tables 9, 10, and 11.

Query C1: Number of transactions on weekends.

Query C2: Number of transactions where amount is over \$10,000.

Query C3: Number of transactions on weekends AND where amount is over \$10,000.

Again, we can see from Tables 9–11 that with ϵ -DP the added noise for the three queries can be multiple times larger than the original values, causing significant information loss. On the other hand, the proposed (ϵ, p') -DP approach yields much more useful results. It appears that $r = 2$ or so is a reasonable choice, which results in at least 94% of the patients ensured with ϵ -DP, together with much more accurate query output.

Overall, for all nine queries on the three databases, as r in the proposed (ϵ, p') -DP increases, the proportion of the individuals ensured with ϵ -DP increases, but the range of confidence interval and the variance of added noise increases at a pace much faster than the proportion of the individuals ensured with ϵ -DP. In summary, while larger r values ensure stronger privacy, smaller r values result in smaller loss in output accuracy, as well as accelerated reduction in the variance of added noise.

5 CONCLUSION

Differential privacy has become one of the most rigorous and commonly applied approaches for answering database queries with controlled privacy disclosure risk. Our study investigated how differential privacy principle can be adapted for database queries when multiple records in a database are associated with the same individual. We have shown that the standard differential privacy mechanism may result in significant information loss when

multiple records per person exist in a database. We have proposed a new privacy principle, (ϵ, p^r) -DP, which incorporates the well-known Pareto principle for analyzing privacy disclosure risks and data utility. With (ϵ, p^r) -DP, information loss in query output can be significantly reduced with a relatively small relaxation in ϵ -DP.

We have used the Pareto principle to implement our proposed approach. However, the basic idea of our approach does not necessarily require the data to follow exactly a Pareto distribution. The proposed approach should be helpful as long as the frequency distribution exhibits an L-shaped curve, like those shown in Figures 3,4, and 5, no matter if it is a Pareto distribution or not. When the actual distribution cannot be well described by a known Pareto distribution, the p value estimated from the data may not correspond to the p parameter in a Pareto distribution. Consequently, the actual proportion of recovery in information loss is likely to be different from the one estimated based on the Pareto distribution. However, the calculated p value is still a reasonable reference value for separating the high-frequency and low-frequency groups. Once the p value is determined, p^r still represents the proportion of individuals with privacy protection not guaranteed with ϵ -DP, regardless of the data distribution. The data owner can control the risk of these high-frequency individuals by choosing an appropriate r value. On the other hand, if the differences in frequency are distributed rather evenly, the standard differential privacy should work well and there is no need to use the proposed approach.

The proposed model has two parameters related to Pareto distribution. The first parameter is p , which is computed from the data. Ideally, p should be much smaller than 0.5 (given the popular “80/20” rule, we expect p to be around 0.2 or smaller). The second parameter is r , which is determined by the user. The user can control the tradeoff between disclosure risk and data utility by specifying an appropriate r value. When p is not small enough, the user can specify a larger r value to get a larger p^r value. Because with the proposed model the proportion of individuals satisfying ϵ -differential privacy is at least $1 - p^r$, by using a larger r the user can still ensure ϵ -differential privacy for the vast majority even when p is not small enough. However, a larger r value will increase information loss since the reduction in information loss is proportional to $(1 - p)^r$. Given a p value computed from the data, the user can try different r values to examine different tradeoff scenarios and choose an appropriate r value to use.

In this study, we have focused our work on count query only. The proposed (ϵ, p^r) -DP model can be extended to other database queries such as the sum and average queries. Continuing on the patient database example in Section 1, assume that the database includes an attribute of the amount charged for each patient visit and a query is sent to find the average cost per patient. Suppose there is a small portion of patients having very frequent visits, each visit associated with a very large amount. Applying standard differential privacy, the added noise will be extremely large, causing perturbed output to be practically useless. In this case, the proposed (ϵ, p^r) -DP model, as specified in Definition 5, can still be applied to separate those few high-cost patients from the vast majority of low-cost patients. The detailed procedure to achieve (ϵ, p^r) -DP in this case requires some context-specific analysis and respective changes in mechanism design. We plan to extend our proposed approach to these different

database queries and other more sophisticated data analysis tasks, which will be interesting and challenging for future research.

Acknowledgments

Xiao-Bai Li's research was supported in part by the National Library of Medicine of the National Institutes of Health under Grant No. R01LM010942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

REFERENCES

- Aggarwal CC and Yu PS (Eds.). 2008 Privacy-Preserving Data Mining: Models and Algorithms. Springer, New York.
- Apple Inc. 2016 Apple previews iOS 10, the biggest iOS release Ever. Apple Newsroom, 6 13, 2017. Retrieved on June 22, 2018 from <https://www.apple.com/bw/newsroom/2016/06/13Apple-Previews-iOS-10-The-Biggest-iOS-Release-Ever>.
- Apple Inc. 2017 Differential Privacy. 11 2, 2017. Retrieved on December 2, 2018 from https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.
- Arnold BC. 2015 Pareto Distribution. Wiley, New York.
- Berndt DJ, McCart JA, Finch DK, and Luther SL. 2015 A case study of data quality in text mining clinical progress notes. ACM Transactions on Management Information Systems 6, 1 (2015), Article 1, 21 pages.
- Dankar FK and El Emam K. 2012 The application of differential privacy to health data. In Proceedings of the 5th International Workshop on Privacy and Anonymity in the Information Society ACM Press, New York, 158–166.
- Dwork C. 2006 Differential privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP'06) 1–12.
- Dwork C. 2011 A firm foundation for private data analysis. Communications of the ACM 54, 1 (2011), 86–95.
- El Emam KK, Dankar FK, Neisa A, and Jonker E. 2013 Evaluating the risk of patient re-identification from adverse drug event reports. BMC Medical Informatics and Decision Making 13, 1 (2013), 114. [PubMed: 24094134]
- Erlingsson Ú, Pihur V, and Korolova A. 2014 RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security ACM Press, New York, 1054–1067.
- European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). 1999 PKDD'99 Discovery Challenge. Retrieved on 3 19, 2017 from <http://lisp.vse.cz/pkdd99/Challenge/chall.htm>.
- Fu Y, Chen Z, Koru G, and Gangopadhyay A. 2010 A privacy protection technique for publishing data mining models and research data. ACM Transactions on Management Information Systems 1, 1 (2010), Article 7, 20 pages.
- Hardy M. 2010 Pareto's Law. The Mathematical Intelligencer 32, 3 (2010), 38–43.
- Harper FM and Konstan JA. 2016 The movielens datasets: History and context. ACM Transactions on Interactive Intelligent Systems 5, 4 (2016), Article 19.
- INFORMS Data Mining Section. 2008 INFORMS Data Mining Contest 2008. Retrieved on 2 9, 2015 from <https://sites.google.com/site/informsdataminingcontest/>.
- Lee J and Clifton C. 2011 How much is enough? Choosing ϵ for differential privacy. In International Conference on Information Security Springer, Berlin, 325–340.
- Machanavajjhala A, Kifer D, Abowd J, Gehrke J, and Vilhuber L. 2008 Privacy: Theory meets practice on the map. In Proceedings of the 24th IEEE International Conference on Data Engineering IEEE Computer Society, Washington, DC, 277–286.
- Sweeney L. 2002 k-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 5 (2002), 557–570.

- Tuarob S, Strong R, Chandra A, and Tucker CS. 2018 Discovering discontinuity in big financial transaction data. *ACM Transactions on Management Information Systems* 9, 1 (2018), Article 3, 26 pages.
- Uber Inc. 2017 Uber/SQL-Differential-privacy. Open Source Software at [Uber.github.io](https://github.com/uber/sql-differential-privacy), 5 1, 2018. Retrived on June 22, 2018 from <https://github.com/uber/sql-differential-privacy>.
- U.S. Census Bureau. 2017 The modernization of statistical disclosure limitation at the U.S. Census Bureau. Meetings, 5 1, 2018. Retrieved on June 22, 2018 from <https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf>.

Fig. 1.
Different distributions for number of occurrences per individual.

Fig. 2.
An illustrative example for (ϵ, p) -differential privacy.

Fig. 3.
Distribution of individuals' occurrences in patient data.

Fig. 4. Distribution of individuals' occurrences in movie data.

Fig. 5.
Distribution of individuals' occurrences in bank data.

Table 1.

Output Ranges of the Query for Illustrative Examples

Database	True Answer to Count Query	Output Mechanism (True Answer + Noise)	Output Range to Query Answer
Figure 1(a)	19	19 + Laplace (5)	(−0.55, 38.55)
Figure 1(b)	38	38 + Laplace (5)	(18.45, 57.55)
Figure 1(c)	100	100 + Laplace (5)	(80.45, 119.55)

Table 2.Tradeoff between ϵ -DP Protection and Accuracy Recovery for Figure 2 Example

r	p^r	Proportion Ensured by ϵ -DP	Proportion of Recovered Loss in Accuracy
0.5	$0.2^{0.5}$	$1 - 0.2^{0.5} = 55\%$	$(1 - 0.2)^{0.5} = 89\%$
1.0	0.2^1	$1 - 0.2^1 = 80\%$	$(1 - 0.2)^1 = 80\%$
2.0	0.2^2	$1 - 0.2^2 = 96\%$	$(1 - 0.2)^2 = 64\%$
3.0	0.2^3	$1 - 0.2^3 = 99\%$	$(1 - 0.2)^3 = 49\%$

Table 3.
Comparison between ϵ -DP and (ϵ, p') -DP on Patient Data for Query A1

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP With $p = 0.1037$ and $r =$	0.5	67.79	6	$29,348 \pm 964$	438
	1.0	89.63	9	$29,348 \pm 2,266$	985
	1.5	96.66	16	$29,348 \pm 7,163$	3,114
	2.0	98.92	21	$29,348 \pm 12,339$	5,365
	2.5	99.65	27	$29,348 \pm 20,398$	8,869
ϵ -DP		100.00	42	$29,348 \pm 49,357$	21,460

Table 4.

Comparison between ϵ -DP and (ϵ, p') -DP on Patient Data for Query A2

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.1044$ and $r =$	0.5	67.69	4	$36,701 \pm 448$	195
	1.0	89.56	7	$36,701 \pm 1,371$	596
	1.5	96.63	13	$36,701 \pm 4,729$	2,056
	2.0	98.91	19	$36,701 \pm 10,101$	4,392
	2.5	99.65	25	$36,701 \pm 17,488$	7,603
ϵ -DP		100	49	$36,701 \pm 67,180$	29,209

Table 5.Comparison between ϵ -DP and (ϵ, p') -DP on Patient Data for Query A3

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.1594$ and $r =$	0.5	60.08	5	$13,921 \pm 700$	304
	1.0	84.06	10	$13,921 \pm 2,798$	1,217
	1.5	93.64	15	$13,921 \pm 6,296$	2,737
	2.0	97.46	21	$13,921 \pm 12,339$	5,365
	2.5	98.99	26	$13,921 \pm 18,915$	8,224
ϵ -DP		100	39	$13,921 \pm 42,558$	18,503

Table 6.Comparison between ϵ -DP and (ϵ, p') -DP on Movie Data for Query B1

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.1327$ and $r =$	0.5	63.57	40	$39,895 \pm 6,083$	2,645
	1.0	86.73	85	$39,895 \pm 27,467$	11,942
	1.5	95.17	128	$39,895 \pm 62,286$	27,081
	2.0	98.24	165	$39,895 \pm 103,500$	45,000
	2.5	99.36	213	$39,895 \pm 172,477$	74,990
ϵ -DP		100	410	$39,895 \pm 639,058$	277,851

Table 7.
Comparison between ϵ -DP and (ϵ, p') -DP on Movie Data for Query B2

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.1575$ and $r =$	0.5	60.31	19	$21,201 \pm 1,372$	597
	1.0	84.25	41	$21,201 \pm 6,391$	2,779
	1.5	93.75	68	$21,201 \pm 17,579$	7,643
	2.0	97.52	93	$21,201 \pm 32,880$	14,296
	2.5	99.02	134	$21,201 \pm 68,262$	29,679
ϵ -DP		100	172	$21,201 \pm 112,468$	48,899

Table 8.
Comparison between ϵ -DP and (ϵ, p') -DP on Movie Data for Query B3

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.1494$ and $r =$	0.5	61.35	10	$9,997 \pm 380$	165
	1.0	85.06	21	$9,997 \pm 1,677$	729
	1.5	94.23	32	$9,997 \pm 3,893$	1,693
	2.0	97.77	50	$9,997 \pm 9,504$	4,132
	2.5	99.14	66	$9,997 \pm 16,560$	7,200
ϵ -DP		100	90	$9,997 \pm 30,793$	13,388

Table 9.Comparison between ϵ -DP and (ϵ, p') -DP on Bank Data for Query C1

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.1973$ and $r =$	0.5	55.59	19	$79,333 \pm 1,372$	597
	1.0	80.27	34	$79,333 \pm 4,395$	1,911
	1.5	91.24	44	$79,333 \pm 7,360$	3,200
	2.0	96.11	56	$79,333 \pm 11,922$	5,183
	2.5	98.27	68	$79,333 \pm 17,579$	7,643
ϵ -DP		100	100	$79,333 \pm 38,017$	16,529

Table 10.
Comparison between ϵ -DP and (ϵ, p') -DP on Bank Data for Query C2

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.2455$ and $r =$	0.5	50.45	37	$21,685 \pm 5,204$	2,263
	1.0	75.45	61	$21,685 \pm 14,146$	6,150
	1.5	87.84	68	$21,685 \pm 17,579$	7,643
	2.0	93.97	80	$21,685 \pm 24,331$	10,579
	2.5	97.01	122	$21,685 \pm 56,584$	24,602
ϵ -DP		100	138	$21,685 \pm 72,399$	31,478

Table 11.Comparison between ϵ -DP and (ϵ, p') -DP on Bank Data for Query C3

Mechanism		Proportion of ϵ -DP Ensured	$f_{p,r}$	True Value $\pm$ 95% Confidence Interval	Variance of Noise
(ϵ, p') -DP with $p = 0.2294$ and $r =$	0.5	52.11	11	$6,285 \pm 460$	200
	1.0	77.06	18	$6,285 \pm 1,232$	536
	1.5	89.01	20	$6,285 \pm 1,521$	661
	2.0	94.74	25	$6,285 \pm 2,376$	1,033
	2.5	97.48	35	$6,285 \pm 4,657$	2,025
ϵ -DP		100	41	$6,285 \pm 6,391$	2,779