UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

26-28 September 2023, Wiesbaden

Differential privacy for microdata

Thijs Benschop (World Bank)

tbenschop@worldbank.org

Abstract

k-anonymity is the standard privacy model for protecting confidentiality in microdata. However, k-anonymity has its limitations, which are partially overcome by extensions, such as l-diversity and t-closeness. The alternative privacy model of differential privacy (DP) has received increasingly more attention in official statistics. For instance, with the US Census Bureau adopting DP. However, DP is primarily designed for protecting database queries and not for releasing microdata. Database queries are not feasible in all contexts either from the data provider's or the data users' point of view and microdata need to be released. In this paper we explore the use of the differential privacy model for disseminating microdata. First, we compare the privacy models k-anonymity and differential privacy. Next, we elaborate on the limitations of the k-anonymity model in practice and describe some applications of differential privacy for microdata. Finally, we conclude by discussing the feasibility of DP as a privacy model for releasing microdata.

Keywords

Differential privacy, k-anonymity, microdata anonymization.

Disclaimer

1 Introduction

k-anonymity is the common privacy model for protecting confidentiality in microdata, such as the release of household survey data. However, k-anonymity has its limitations, which are partially overcome by extensions, such as ℓ-diversity and t-closeness. Dwork (2006) introduced an alternative privacy model: differential privacy. Differential privacy (DP) has received increasingly more attention in official statistics. For instance, with the US Census Bureau adopting DP for the 2020 Census. Often, differential privacy is considered superior to k-anonymity, as it provides a formal privacy guarantee. However, differential privacy is primarily designed for protecting database queries and not easily applicable to protecting microdata for dissemination. Database queries are not feasible in all contexts either from the data provider's (high cost of maintaining a system for dynamic queries) or the data users' (less flexibility) point of view and microdata need to be released. Microdata are datasets that provide information on a set of variables for each individual respondent. Respondents can be natural persons, but also legal entities such as companies. Microdata provide more flexibility to users of the data than tables and aggregated statistics. Section 2 of this paper describes the privacy models k-anonymity and differential privacy. Section 3 elaborates on some of the limitations of k-anonymity and section 4 gives an overview of some implementations of differential privacy for the release of microdata. Finally, section 5 concludes and describes some ideas for future work.

2 Privacy models

This section defines two formal privacy models: k-anonymity and differential privacy. k-anonymity is a syntactic model, where the data is considered safe once a syntactic condition is met. Differential privacy provides a formal mathematical guarantee of privacy protection.

2.1 k-anonymity

The concept of k-anonymity was first proposed by Samarati and Sweeney (1998). They noticed that removing explicit or direct identifiers, such as names, addresses, phone numbers or Social Security Numbers is not sufficient to protect the identities of the records in microdata. Often, other attributes in the dataset, which are individually not directly identifying the records, render the records identifiable when combined with other attributes and linked to another external dataset which includes direct identifiers. For instance, 87% of the US population can be uniquely identified by only using gender, date of birth and the ZIP code (Sweeney, 2000). The attributes that are used to identify the records are called quasi-identifiers.

The identification process is illustrated in Figure 1. The survey dataset on the right does not include direct identifiers. However, using the external dataset on the left, the first record can be successfully identified by matching on the variables sex and date of birth. The attributes sex and date of birth are quasi-identifiers. This allows to reveal the income of Joe A., which is considered sensitive information and hence this constitutes a breach of confidentiality.

External dataset

Name	*Sex*	*Date of birth*
Joe A.	M	Dec 1963
Jane B.	F	Feb 1968

Survey dataset

Name	*Sex*	*Date of birth*	*Income*
.	M	Dec 1963	80,000
.	M	Jul 1963	60,000
.	F	Feb 1968	25,000
.	F	Feb 1968	150,000

Figure 1 Illustration of record mathing

A dataset fulfils k-anonymity, if each combination of quasi-identifiers (called a key), occurs at least k times in the dataset. This guarantees that an intruder faces uncertainty to which record to match using the set of quasi-identifiers, as the intruder will find at least k candidates. If each record is part of a larger group, then any of the records in this group could correspond to a single individual.

If a dataset does not fulfil k-anonymity, generalization and suppressions can be used to achieve k-anonymity. Generalization means that some levels of the quasi-identifiers are combined. This leads to a decrease of the number of possible combinations/keys and an increase of the number of records sharing the same key. An example of a generalization technique is recoding. Figure 2 illustrates the generalization of the variable *Date of Birth*, where all birth months in the same year are grouped together. Now it is no longer certain which of the two records Joe A. matches with.

Survey dataset after generalization

Name	*Sex*	*Date of birth*	*Income*
.	M	1963	80,000
.	M	1963	60,000
.	F	1968	25,000
.	F	1968	150,000

Figure 2 Illustration of generalization

The definition of k-anonymity depends on the selection of quasi-identifiers. Attributes that can be used by an intruder for matching with an external dataset are identified as quasi-identifiers. This requires assumptions on the knowledge of an intruder and external data a potential intruder has for performing a matching attack.

2.2 Differential privacy

In differential privacy, privacy protection is a property of the data processing method, rather than the dataset itself. Therefore, differential privacy is designed for protecting interactive queries from a database and is not directly suited for anonymizing microdata to be released. Unlike k-anonymity, differential privacy does not depend on the selection of quasi-identifiers and therefore no assumptions are needed on the intruders and the knowledge the intruders have. So even if the intruder learns new information or gets access to new methods to identify records, differential privacy guarantees the privacy of the records rendering the method future-proof.

Informally, differential privacy guarantees that the outcome of a query does not change significantly if one record is removed from or added to the dataset. If this is the case, an intruder cannot infer new information on a particular record from the queries. Formally, differential privacy is defined as:

Definition ε-differential privacy: Assume a mechanism $\mathcal{A}$ that randomizes query outputs and any pair of neighbouring databases $\mathcal{D}$ and $\mathcal{D}'$. Then, $\mathcal{A}$ satisfies ε-differential privacy iff

$$P[\mathcal{A}(\mathcal{D}) = \mathcal{S}] \leq \exp(\varepsilon) * P[\mathcal{A}(\mathcal{D}') = \mathcal{S}]$$

where $\mathcal{S} \in$ Range($\mathcal{A}$). $\mathcal{D}$ and $\mathcal{D}'$ are neighbouring databases if they differ in exactly one record, i.e., $\mathcal{D}'$ is generated by removing or adding exactly one record to or from $\mathcal{D}$. ε is called the privacy budget and is set by the user. The larger the privacy budget (the higher ε), the lower the privacy protection.

To satisfy differential privacy, uncertainty needs to be added to the query result. The most common method applied is the addition of noise through a Laplace mechanism. Noise is added to the data to mask the contribution of any possible record to the query result. The amount of noise that needs to be added depends on the available privacy budget ε as well as the sensitivity of the query function. The sensitivity of a function generating the query result from the database is the maximal change caused to the query result when one record is added or removed and is defined as

Definition sensitivity: $\Delta f = \max_{\mathcal{D},\mathcal{D}'} ||f(\mathcal{D}) - f(\mathcal{D}')||$, where f is the function generating the query results.

The larger the sensitivity, i.e., the more one single record contributes to the function f, the more noise needs to be added to the data. The Laplace mechanism satisfies **ε**-differential privacy if the random noise added to the query result is sampled from a Laplace distribution with mean μ=0 and scale b=Δf/ε.

A simple example is the sensitivity function for the count of the number of females in a dataset. The sensitivity of this query function is 1 (the count can change at most one, when one record is added or removed from the dataset) and hence the noise to be added to the query result is drawn from a Laplace distribution with mean μ=0 and scale b=1/ε, where the privacy budget ε should be set by the user. Here it should be noted that to respect the privacy budget ε, no additional questions can be made to the dataset, as the full privacy budget is used by this query.

In a simple example as a count query, the sensitivity function is straightforward. However, if the query output is a full microdata dataset, the sensitivity is much more complex and larger, which hence requires much higher levels of noise to be added. The addition of noise to protect the privacy of records in the data results in a loss of data utility. Noise addition makes the data less accurate or useful for certain types of analysis.

3 Challenges to k-anonymity in practice

k-anonymity has drawn various criticism as to the extent it protects the data and poses challenges applying it in practice to microdata. In this section we elaborate on some of the criticism and challenges as well as solutions.

k-anonymity is highly dependent on the selection of quasi-identifiers. Only those variables selected as quasi-identifiers are protected from being used in a matching attack. The selection of quasi-identifiers is based on assumptions of the variables available in the external datasets available to a potential intruder. Often, not all external datasets are known to the data producer, e.g., private datasets, and new datasets may appear in the future. This may lead to an inadequate protection of the data.

Linked to this is also the limitation to the number of quasi-identifiers selected. When k-anonymity is applied to many quasi-identifiers, the need for generalization and suppression is typically very large, which has a detrimental impact on the data utility. In practice, in a household survey dataset, the number of quasi-identifiers selected should not exceed 10-15. However, a dataset may contain more quasi-identifiers.

In samples with low sample proportion, k-anonymity may lead to overprotection as the frequencies of each key is low due to the size of the sample. This may lead to overprotection, which can be seen in the following example. Assume a 5 percent sample of a population of 5 million. The sample size is 250,000. Assume we have achieved 3-anonymity in this sample by applying generalization methods. Subsequently, we draw a 20 percent sample of this sample (being effectively a 1 percent sample of the full population). It is likely that this sample will not satisfy 3-anonimity, as we have dropped 80 percent of the records from the first sample. Further generalization or suppression will be needed to achieve 3-anonymity. However, this likely overprotects the second sample, as the second 1 percent sample is obviously less risky than the first 5 percent sample. The smaller sample contains fewer records and information. This is caused by the fact that k-anonymity is applied to the sample, rather than the population, as the full population is unknown.

Often generalization techniques, such as recoding or microaggregation, are not sufficient to achieve k-anonymity for a given threshold k and suppression methods are applied. Suppression methods induce missing values in the microdata and increase the frequency counts of keys by interpreting missing values as any of the value in the range of the given attribute. For example, if some values in the variable region with values regions A, B, C, D and E are suppressed, the missing values are assumed to be any of the five regions when counting the frequencies of the keys. This is illustrated in Figure 3. This seems an unrealistic assumption, as the fifth record can only

assume one region in reality, not five simultaneously. Therefore, the frequencies may be overestimated and the data underprotected. A partial solution to this is to count a match with a missing value only partially.

Region	Age	Gender	Frequency
A	30-39	F	1
B	30-39	F	1
C	30-39	F	1
D	30-39	F	1
E	30-39	F	1

Region	Age	Gender	Frequency
A	30-39	F	2
B	30-39	F	2
C	30-39	F	2
D	30-39	F	2
.	30-39	F	5

Figure 3 Illustration of effect of missing values in k-anonymity

Furthermore, to select the values that need to be suppressed to achieve k-anonymity, algorithms are used to reduce the number of suppressions. However, these algorithms are more likely to suppress values with low frequencies/values in rare categories, which can provide the intruder with a clue how to reengineer the suppressed values.

k-anonymity guarantees that the intruder finds at least k matches when matching on the set of quasi-identifiers. However, this may not protect the information in sensitive variables in the data sufficiently. Sensitive variables contain information that cannot be used for matching but may inflict harm on the respondent when the information is revealed. Examples of sensitive variables are variables related to health, income, or religion. If there is no variation in the values of the sensitive variable within the records that share the same key, the intruder can still learn the sensitive attribute without the need for exact matching. In the example in Figure 4, the intruder cannot match exactly on the quasi-identifiers region, age and gender. Yet, the intruder can learn the sensitive health condition attribute for a person living in region A, who is in the age range of 30-39 and female, since the value is *Yes* for all five records that share this key.

Region	Age	Gender	Health condition	Frequency
A	30-39	F	Yes	5
A	30-39	F	Yes	5
A	30-39	F	Yes	5
A	30-39	F	Yes	5
A	30-39	F	Yes	5

Figure 4 Example of disclosure without exact matching

ℓ-diversity, a measure proposed by Machanavajjhala et al. (2007) as a solution to the above-described problem, is an extension to k-anonymity. ℓ-diversity protects data against attribute disclosure of sensitive attributes by ensuring that each sensitive attribute has at least ℓ "well represented" values in the group of records sharing the same key. However, ℓ-diversity is limited by the number of different values in the sensitive attribute. Therefore, in the example above, where the sensitive attribute Health condition can only assume the values yes or no, the maximum level of ℓ can be 2.

k-anonymity is defined for categorical quasi-identifiers. Often, quasi-identifiers are both categorical and continuous variables, e.g., age or income. To use k-anonymity for these continuous quasi-identifiers, they need to be converted into a categorical variable, e.g., by using microaggregation for income or constructing intervals for age. This implies a large loss of utility for the user, as many types of analyses depend on the availability of continuous variables or the results of the analysis will be very different, e.g., regression.

Finally, the threshold for k-anonymity to be used must be set. In the literature we find several standard thresholds, such as 3 or 5, but in other studies much higher thresholds are used. In medical studies the threshold for k can be as large as 50. The threshold should be defined as a function of the willingness to take risk. The higher the threshold, the higher the level of generalization and suppression required, the higher the loss of utility of the data for the user. To some extent, the selection of the threshold to be used is arbitrary and follows rules of thumb but lacks a theoretical underpinning. This may lead to underprotection at the cost of confidentiality, or overprotection at the cost of data utility.

4 Differential privacy for microdata

As noted above, differential privacy is a property of the data processing method, rather than the dataset itself. When applying differential privacy to microdata, the full dataset should be considered the output of the DP algorithm and hence none of the records in the dataset should have a significant impact on the released dataset. No single record from the original dataset can be identifiable in the released dataset. This can only be achieved by adding large levels of noise to the records in the dataset. The noise results in a loss of utility of the data for the users. There have been several implementations of differential privacy for microdata and in this section we discuss some w.r.t. the mechanism used to achieve differential privacy as well as the information loss.

Lee and Chung (2020) introduce Informative attribute Preserving (IPA) for protecting medical microdata. IPA uses generalization as well as suppression to reduce the amount of noise that needs to be added to the data. Furthermore, to reduce the number of variables to treat, they distinguish between dimension attributes (similar to quasi-identifiers in k-anonymity) and informative attributes (similar to sensitive variables), that are considered unknown to the intruder and can hence not be used for identification. After generalization, the dimension attributes are suppressed for equivalence classes with fewer records than a stochastic threshold t. The threshold

is set at t, but Laplace noise is added to the threshold, as otherwise this threshold contains information for the intruder. Subsequently, counterfeit records, which are records that are made up, are added to some equivalence classes. Again, a Laplace mechanism is used to select the equivalence classes to which counterfeit records are added as well as to determine the number of records to be added. The exponential mechanism is used to determine the values of the informative attributes of the counterfeit records. Finally, the above steps are used to generate different generalizations and one option is selected from the set of possible outcomes. All four steps are differentially private and hence, by the sequential composition theorem, so is IPA.

Lee and Chung (2000) used IPA on a medical dataset with 1,361,000 records, five dimension attributes and one information attribute to show that the information loss for achieving differential privacy with $\varepsilon = 1$ is considerably lower than the information loss caused by using generalization and suppression to achieve 10-anonimity, while differential privacy arguably protects the data better than k-anonymity.

Muralidhar et al. (2020) use two approaches to generate differentially private microdata: 1) differentially private synthetic microdata from noise-added covariates and 2) noise addition to the cumulative distribution function. The synthetic data are generated by sampling from a multivariate normal distribution, where the mean vector and covariance matrix are differentially private. To obtain the DP versions of the mean vector and covariance matrix, Laplace noise is added to the sum of each attribute, the sum of the squared values of each attribute and the sum of the product of each pair of attributes. The privacy budget ε is distributed over these $2m + m(m-1)/2$ sums, where m is the number of attributes in the dataset. Muralidhar et al. (2020) note that this method is only suitable for dataset with few attributes.

For the second method, DP values for the attributes are generated by sampling from a univariate distribution for each attribute with a mean and variance that have been perturbed using Laplace noise. Subsequently, rank swapping is used, where each original value is swapped with the sampled value with the same rank within the same attribute.

Muralidhar et al. (2020) compare both methods on data utility and data protection, where the utility is assessed by evaluating the correlation structure and data protection by measuring the difficulty of correctly matching records. They use a privacy budget ε=1 and find that utility and protection levels that are very different across methods and find that the protection is not dependent on ε, but rather on the method applied. The more noise is added to the data, the higher the protection. Therefore, they conclude that differential privacy is not suitable for application to microdata.

5 Conclusion

Even though k-anonymity has its shortcomings in terms of providing a formal privacy guarantee and the implementation poses some challenges for anonymizing microdata, such as household survey data, it remains the mostly used privacy model for releasing microdata. The reasons for most practitioners to stick with k-anonymity for microdata are manifold. The alternative model of differential privacy is not well-suited to protect microdata: the few studies that applied differential privacy techniques to microdata have generally shown that the amount of

noise that needs to be added is very high, leading to large losses of utility, or that the methods are not suited for dataset with many key variables. Suggested solutions to these problems are to apply differential privacy only to a selected set of quasi-identifiers or combining noise with generalization and suppression techniques. However, this approach introduces the drawbacks of these methods and reduces the strength of the privacy guarantee for which differential privacy is appealing. Differential privacy has also its own limitations. One question is how to set the privacy budget ε for differential privacy. Unlike the parameter k for k-anonymity, the interpretation of ε does not relate in a straightforward way to the probability of re-identification of a record by an intruder. Differential privacy relies on the introduction of noise, which can have large impacts in the utility of microdata. For example, noise addition can introduce unlikely combinations in a dataset, such as a 5 year old that has a university degree. Many practitioners shy away from using perturbative methods that rely on noise or this reason and prefer generalization and suppression methods that have more predictive outcomes in terms of utility. Finally, several studies have shown that the differences in confidentiality protection and utility loss depends largely on the data and less on the formal privacy model and methods applied. In conclusion, differential privacy is a powerful privacy model for protecting query outputs, but k-anonymity remains more suitable for protecting microdata for release.

References

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S. Kifer, D., Leclerc, P., Sexton, W., Simpson, A., Task, C. and Zhuravlev, P. (2021). *An Uncertainty Principle is a Price of Privacy-Preserving Microdata*. https://arxiv.org/abs/2110.13239

Dwork, C. (2006, July). *Differential privacy*. In International colloquium on automata, languages, and programming (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg.

Khan, M.I., Foley, S.N. and O'Sullivan, B. (2021*). From k-anonymity to Differential Privacy: A Brief Introduction to Formal Privacy Models*. ⟨hal-03226881⟩

Lee, H. and Chung Y.D. (2020). *Differentially private release of medical microdata: an efficient and practical approach for preserving informative attribute values.* BMC Medical Informatics and Decision Making 2020 20:155.

Machanavajjhala, A., Kifer, D. , Gehrke, J. and Venkitasubramaniam, M. (2007*). ℓ-diversity: Privacy beyond k-anonymity*. ACM Trans. Knowl. Discov. Data, 1(1):3, 2007.

Muralidhar, K., Domingo-Ferrer, J. and Martínez, S. (2020). *ϵ-Differential Privacy for Microdata Releases Does Not Guarantee Confidentiality (Let Alone Utility).* 10.1007/978-3-030-57521-2_2.

Samartati, P. and Sweeney, L. (1998). k-anonymity: a model for protecting privacy. *Proceedings of the IEEE Symposium on Research in Security and Privacy* (S&P). May 1998, Oakland, CA.

Sweeney, L. (2000). *Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4*. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA.