DIFFERENTIAL PRIVACY FOR MICRODATA

Thijs Benschop, World Bank

WORLD BANK GROUP

UNECE Expert Meeting on Statistical Data Confidentiality

26-28 September 2023, Wiesbaden

Outline

1. Two privacy models:
    1. k-anonymity
    2. differential privacy (DP)
2. Challenges to k-anonymity in practice
3. Applications of DP to microdata
4. Conclusion and outlook

k-anonymity

- Microdata dataset fulfills k-anonymity if each combination of keys appears at least k times (see Samarati and Sweeney, 1998)

- Often generalization techniques and suppression are used to achieve k-anonymity

- k-anonymity depends on the definition of quasi-identifiers

Example k-anonymity

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

Differential privacy

- Privacy is property of data processing method in Differential Privacy (DP) (Dwork, 2006)
- Designed to protect data queries -> not microdata
- Differential privacy guarantees that the outcome of a query does not change significantly if one record is removed from or added to the dataset

Differential privacy - definition

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

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

where $\mathcal{S} \in \mathrm{Range}(\mathcal{A})$. $\mathcal{D}$ 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.

- To satisfy DP, uncertainty is added through noise
- The amount of noise depends on the sensitivity and privacy budget

Definition sensitivity: $\Delta f = \max_{\mathcal{D},\mathcal{D}'} ||f(\mathcal{D}) - f(\mathcal{D}')||$,

where f is the function generating the query results.

Challenges to k-anonymity (1)

- No formal privacy guarantee
- Dependent on selection of quasi-identifiers (need to make assumptions/may change in future)
- In practice limit on number of quasi-identifiers
- In case of a low sample proportion, may lead to overprotection
- Interpretation of missing values (introduced by suppression techniques)

Challenges to k-anonymity (2)

- Sensitive variables not always well protected
  -> l-diversity
- Not possible to combine categorical and continuous key variables
- Choice of threshold k

DP implementations for microdata (1)

- How can DP be applied to release microdata?
- What is the DP algorithm and what is its output?
- Microdata dataset itself can be regarded as output
- Need to apply noise to microdata

DP implementations for microdata (2)

- Informative attribute Preserving (IPA) for protecting medical microdata (Lee and Chung, 2020)
- IPA uses generalization as well as suppression to reduce the amount of noise that needs to be added to the data
- Distinction between dimension and informative attributes

DP implementations for microdata (3)

- Lee and Chung (2020) use IPA on medical dataset with five dimension attributes and one information attribute using a privacy budget equal to 1
- Results compared with 10-anonymity: better protection and higher utility

DP implementations for microdata (4)

- Muralidhar et al. (2020) use two approaches to generate differentially private microdata
  1) differentially private synthetic microdata from noise-added covariates
     - sampling from multi-variate normal distribution with DP versions of mean vector and covariance matrix
     - Only suitable for datasets with few variables
  2) noise addition to the cumulative distribution function
     - Sampling from univariate distributions followed by rank swapping

DP implementations for microdata (5)

- Muralidhar et al. (2020) use privacy budget of 1 and compare methods on utility and risk
- Generally, the more noise is added, the lower the risk -> large levels of noise are needed to protect the data
- Conclude that DP is not suitable for microdata protection

Conclusion

- Despite shortcomings k-anonymity remains the most used privacy model for microdata
- Main reason: DP is not suitable for microdata protection
  - Too large levels of noise
  - No implementations for large datasets
  - No clear interpretation of privacy budget
- Need for further improvements of k-anonymity

Thank you for your attention

References

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

- 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.

- 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.