DIT IS DE TITEL VAN MIJN AFSTUDEERVERSLAG

by

Tjibbe van der Ende

in partial fulfillment of the requirements for the degree of

Master of Science
in Software Engineering

at the Open University, faculty of Management, Science and Technology
Master Software Engineering
to be defended publicly on Day Month DD, YYYY at HH:00 PM.

Student number:	student number	
Course code:	IMA0002	
Thesis committee:	titles and name of the chairman (chairman),	Open University
	titles and name of the supervisor (supervisor),	Open University

CONTENTS

1
INTRODUCTION

1.1. RESEARCH QUESTIONS

Main question:
How can the nD-Laplace algorithm be applied in training privacy-preserving clustering algorithms on distributed n-dimensional data?

1. RQ1: How can 2D-Laplace be used to protect the data privacy of 2-dimensional data which is employed for training clustering algorithms?
2. RQ2: How can 3D-Laplace be extended to protect the data privacy of n-dimensional data which is employed for training clustering algorithms?
3. RQ3: What is the impact of different privacy budgets, dataset properties, and other clustering algorithms on the research conducted for research question 2?

2

LITERATURE REVIEW

This chapter lays out the theoretical foundation of this work. To review the past literature, it is first necessary to gather the required knowledge for it.

2.1. Differential privacy

2.1.1. Laplace algorithm

2.2. Clustering

2.2.1. Methods

2.2.2. Evaluation methods

2.3. Literature review

3

nD-Laplace

3.1. 2D-Laplace

The theory for this subject is heavily inspired by the paper that was written by Andrés et al. [Andrés et al., 2012]. This notion of geo-indistinguishability was introduced to solve the issue of privacy and location data. It offers an alternative approach for differential privacy by adding noise to the location locally before sending it to a location-based system (LBS) like Google maps. This section starts with an introduction to mathematics for the planar and polar Laplace algorithm. For each of the different subsections, we visualize and explain open challenges and theoretic for applying them for clustering.

Math symbols

X Set of locations for a user. (R^2).

Z For every $x \in X$ a perturbed location $z \in Z$ is reported..

ϵ Defined as $\epsilon = l/r$.

θ Angle.

l Privacy level.

r Radius.

3.1.1. Planar and polar Laplace

The idea of planar Laplace is to generate an area around $x0 \in X$ according to the multivariate Laplace distribution . The mechanism of planar Laplace is a modification of the Laplace algorithm to support distance [Andrés et al., 2012]. This distance method $dist(x, x')$ is defined as the Euclidean distance between two points or sets. Recalling the definition of Laplace, this method $|x - x'|$ is replaced by the distance metric. Hence, the definition of the Probability Density Function (pdf) by Andrés et al. is:

$$\frac{\epsilon^2}{2 * \pi} e(-\epsilon d(x_0, x)) \tag{3.1}$$

Which is the likelihood a generated point $z \in Z$ is close to $x0$. The method works for Cartesian coordinates but was modified to support polar coordinates by including θ. So each point is reflected as (r, θ) and can be modified by using a slight modification to work for polar Laplace.

A point $z \in Z$ where $z = (r, \theta)$ is randomly generated using two separate methods.

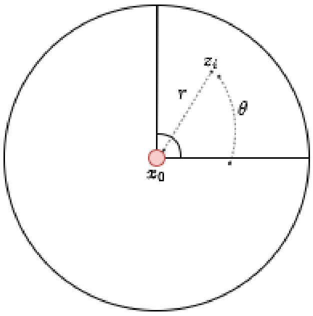

Figure 3.1: Representation of the generated $z = r\theta$ and original point $x0$.

Calculating r: This variable is described as $dist(x_0, z)$ and can be randomly drawn by inverting the CDF (**Link**) for the Laplace distribution:

$$C_\epsilon{}^{-1}(p) = -\frac{1}{\epsilon}(W_-1(\frac{p-1}{e})+1) \tag{3.2}$$

For this equation, W_-1 is a Lambert W function with -1 branch. The Lambert w function, also called the product logarithm is defined as $W(x)e^{W(x)} = x$ [Lehtonen, 2016]. The purpose of the Lambert w function is to invert the CDF of the Laplace distribution to generate random noise for one of the coordinates (r) using the random value of p.

Calculating θ: The other coordinate (θ) is defined as a random number $[0, 2\pi]$.

To visualize these methods it is necessary to convert the polar coordinates for $z = (r, \theta)$ back to a plane (x, y). This is described as step 4 of the planar Laplace algorithm [Andrés et al., 2012] and visualized using figure 1.

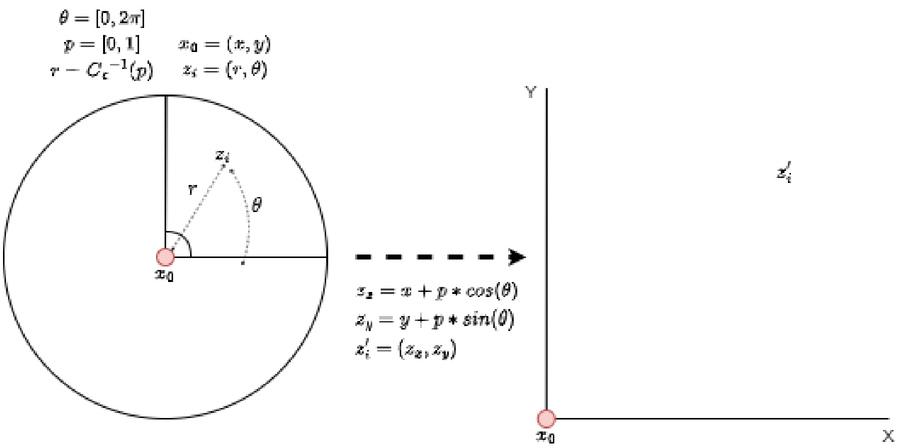

Figure 3.2: Representation of converting the perturbed point $z = (r, \theta)$ to a point z_x, z_y

We highlight the CDF function for assessing if the probability of a random point falls between 0 and r:

Can be moved t appendix

$$C\epsilon(r) = D_\epsilon, R(p)dp = \int_0^r 1-(1+\epsilon r)e^{-\epsilon r} \tag{3.3}$$

3.1.2. Truncation

3.1.3. Optimizing for clustering

A drawn area as shown in **??** can be expressed as a perturbation area P_{area}.

3.2. 3D-Laplace

Is considered for research question 3

4
Methodology

To gain insights into the proposed methods for researching the appliance of (ND)-Laplace for cluster algorithms we conducted experiments. The experiment results are used to evaluate our method against other literature. In this chapter we explain:

1. Datasets
2. Environmental setup.
3. For each research question: Description of the different experiments.
4. For each research question: Results.

4.1. Datasets

For this research, we will use a synthetic dataset for all three research questions.

Records	Centers	Dimensions	Standard deviation	Research
200	4	2	0.60	RQ 1

Research question 3 uses a "real-world" dataset to properly assess the different dataset properties that are the subject of this research question.

Describe datasets

4.2. Environmental setup

Describe the exact environment details

4.3. Experiment setup

4.3.1. Research question 1

4.3.2. Research question 2

4.3.3. Research question 3

4.4. Results

4.4.1. Research question 1

4.4.2. Research question 2

4.4.3. Research question 3

Bibliography

Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *CoRR*, abs/1212.1984, 2012. 6, 7

Jussi Lehtonen. The Lambert W function in ecological and evolutionary models. *Methods in Ecology and Evolution*, 7(9):1110–1118, 2016. ISSN 2041-210X. doi: 10.1111/2041-210X.12568. 7

Math symbols

X Set of locations for a user. (R^2). 6

Z For every $x \in X$ a perturbed location $z \in Z$ is reported.. 6

ϵ Defined as $\epsilon = l/r$. 6

θ Angle. 6

l Privacy level. 6

r Radius. 6