Differential Privacy with δ-Neighbourhood for Spatial and Dynamic Datasets

Chengfang fang
Huawei International,
Singapore
fang.chengfang@huawei.com

Ee-Chien Chang
School of Computing,
National University of Singapore
changec@comp.nus.edu.sg

NUS
National University
of Singapore

1. Introduction

1. Introduction

- Differential Privacy provides a strong guarantee but leads to low utility.

- Under the notion of differential privacy, a dataset is a set of records (each record contributed by an entity) where all records are treated “equally”.

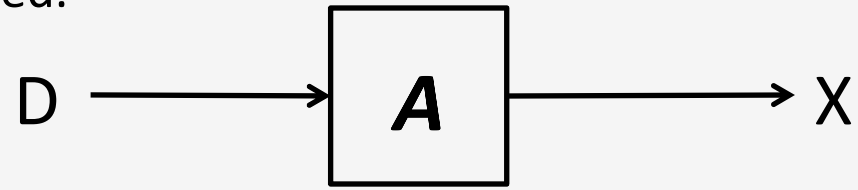
- In some applications, the records have an underlying distance function. Adversary probably have more background information on nearby entities.

- We want to formulate a notion that captures the above intuition, and exploit it for better utility by investing more budget on nearby entities.

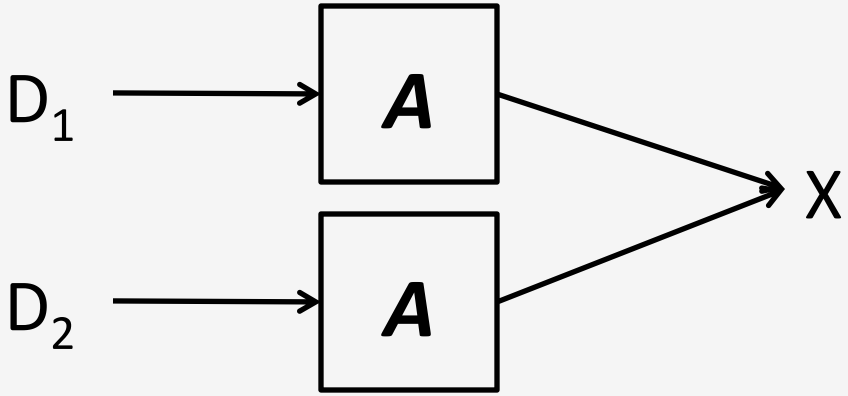
2. Main Idea

Background: Differential Privacy

A mechanism *A* extracts information X from a dataset D. The extracted X is to be published.



The mechanism is differentially private if given an output X, for any two datasets D_1 and D_2 that differ in one entity, the probability that D_1 leads to X is “bounded” by the probability that D_2 leads to X.



A mechanism ***A*** is *ε*-differential private if:

for any X,

$$\Pr(\boldsymbol{A}(D_1) = X) \leq e^{\varepsilon}(\boldsymbol{A}(D_2) = X)$$

for any two “neighbouring” datasets D_1 and D_2 .

Background: Datasets Neighbourhood

There are a few variants of neighborhood adopted in the literatures.

- *Bounded neighbourhood*: D_1, D_2 are neighbours if,
  $D_1 = D_2 - \{ x \}$ for some x. ---(1)

- *Unbounded neighbourhood*: D_1, D_2 are neighbours if,
  $D_1 = D_2 - \{ x \} + \{ y \}$ for some x,y. ---(2)

- *"Combined"*: D_1, D_2 are neighbours if either (1) or (2) holds.

Let us call the above "standard" neighbourhood.

Proposed formulation

We use a variant of neighbourhood derived from an underlying distance function on the entities.

Let the source S be a set of spatial points.
D_1, D_2 are δ-neighbours with respect to the source S if:

$D_1 = D_2 - \{ x \}$ for some x, s in S, where $d(x,s) < \delta$

or

$D_1 = D_2 - \{ x \} + \{ y \}$ for some x, y where $d(x,y) < \delta$

Similarly, we can define **δ-neighbourhood**, **differential privacy under δ-neighbourhood.**

Remarks

The proposed formulation makes very slight but subtle modification of the original. Two issues:

- What is the interpretation of the modification?
- How to exploit the modification for higher utility?

We tackle these issues in this paper/talk.

3. Interpretations

Standard vs δ-neighbourhood

- We can define a distance function between datasets: The distance between D_1 and D_2 is the edit distance (min of insertion, deletion, replacement to get D_2 from D_1). The bounds assured by a d.p mechanism is proportional to such distance.

- Standard and δ-neighbourhood lead to different edit distances.

D_1	D_2	standard	δ=5
{0,0,0,0}	{0,0,1,1}	2	2
{0,0,0,0}	{0,0,0,9}	1	2

edit distance between D_1 and D_2

- Hence, δ-neighbourhood can be viewed as a re-distribution of assurance, which takes into consideration the distances among the entities.

Scenario 1

- Consider an application where there are constrains on the dataset, so that an arbitrary set of points D drawn from an universe U may not be valid (i.e. probability that D occurs is 0). Suppose

  D_1 and D_2 are valid and are neighbours (standard)

  implies

  they are neighbours under δ-neighbourhood.

  then the two notions are equivalent.

For example, D consists of locations of a moving vehicle in a day sampled at every minute. Due to the speed limit of the vehicle, certain sets are not valid.

Scenario 2

From the perspective of a data contributor Bob.

- (Standard Neighbourhood) Bob accepts that there are many entities ***in the world*** with the same background information as him, and thus from the published data, the adversary cannot easily infer whether he is in the dataset D. Hence Bob is willing to contribute.

- (Neighbourhood) Bob accepts that there are many entities, ***within distance δ,*** with the same background information as him, and thus from the published data, the adversary cannot easily infer whether he is in the dataset D. Hence Bob is willing to contribute.

Scenario 3

Suppose D is already published *ε*-differential privately under *δ*-neighbourhood, but Bob is more comfortable to accept that there are many entities ***within distance 2δ*** having the same background information as him.

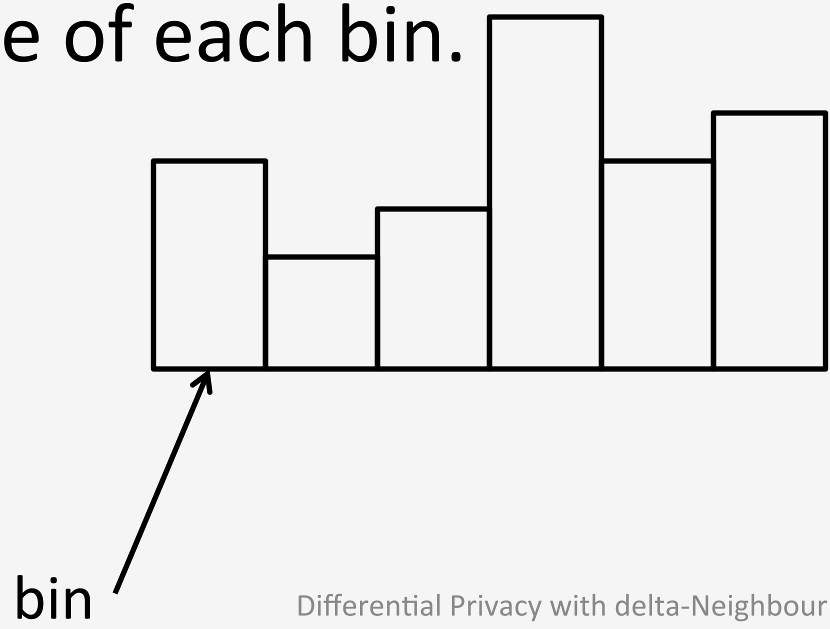
Note that there are still assurance on the bounds, but with weaker bound: ***2ε.***

Hence, *δ*-neighbourhood can be viewed as a redistributed of assurance, where the parameter ***δ*** determines the rate of distribution.

4. Utility

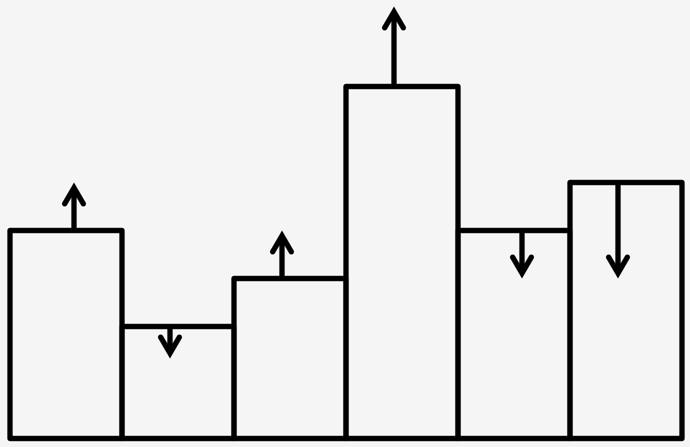
Background: histogram

- Consider a dataset D which is a multi-set of real numbers.
- Histogram represents distribution of D, counting the number of elements within the range of each bin.



Background: publishing histogram d.p.

- A well-known diff. private mechanism adds i.i.d Laplace noise to the bins.



Background: Sensitivity

- A mechanism that publishes

f(D) + Laplace noise

achieves diff priv. The level of noise depends on the *sensitivity* of the function f().

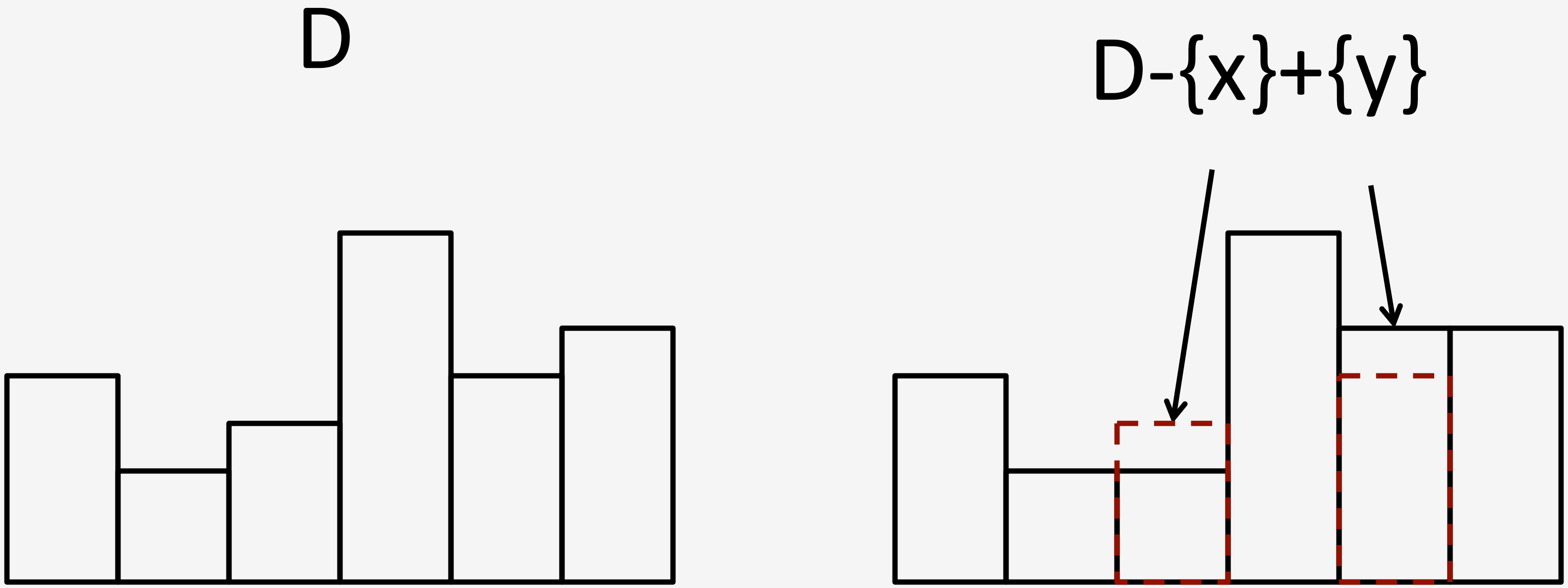
Sensitivity is defined as:

$$\sup || \boldsymbol{f}(D_1) - \boldsymbol{f}(D_2) ||_1$$

where sup is taken over all pairs of neighbours.

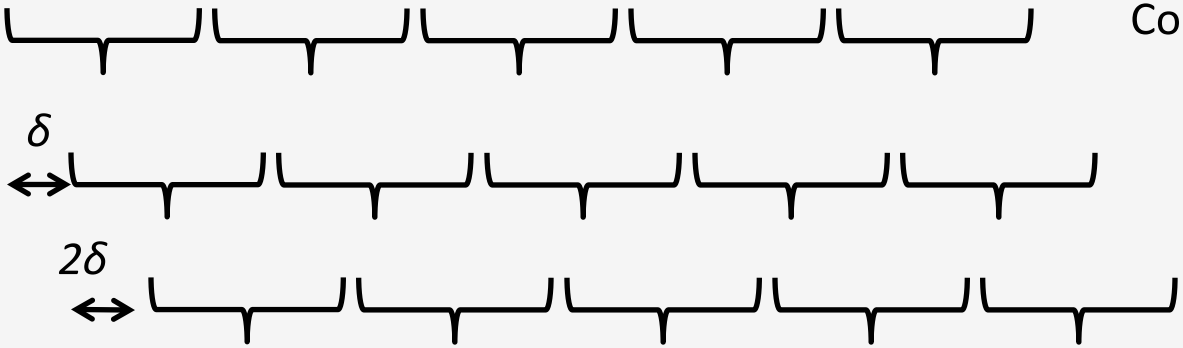
Histogram publishing

- It is well known that sensitivity of publishing histogram is 2, regardless of whether it is under either standard or δ-neighbourhood.



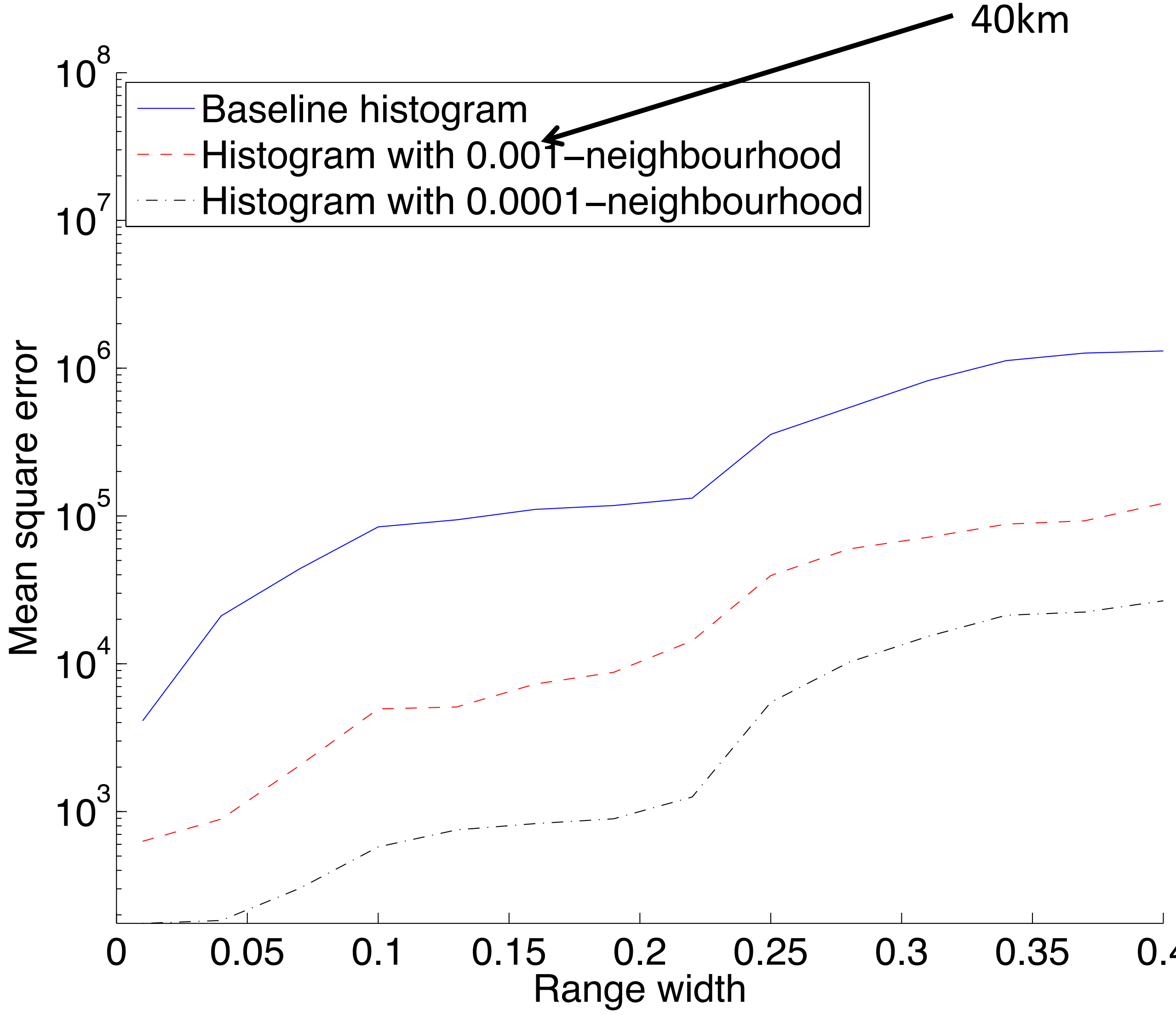
Proposed method (illustration for 1d)

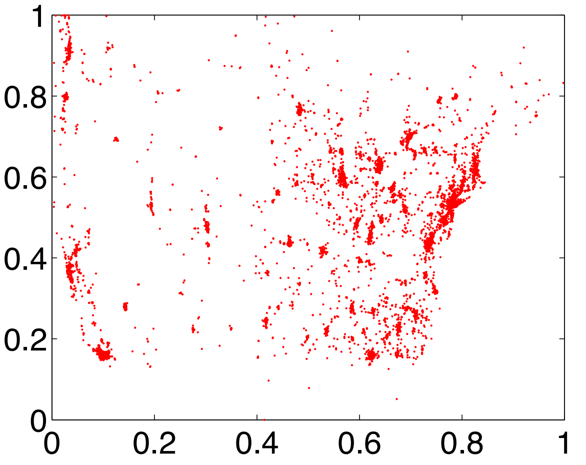
- Instead of publishing one single histogram, publish a series of histograms, shifted by *δ*.



- The sensitivity is still 2, but the utility can be significantly increased, as more information is published.
- We give a general approach (see paper for details).

Experimental result for 2d points

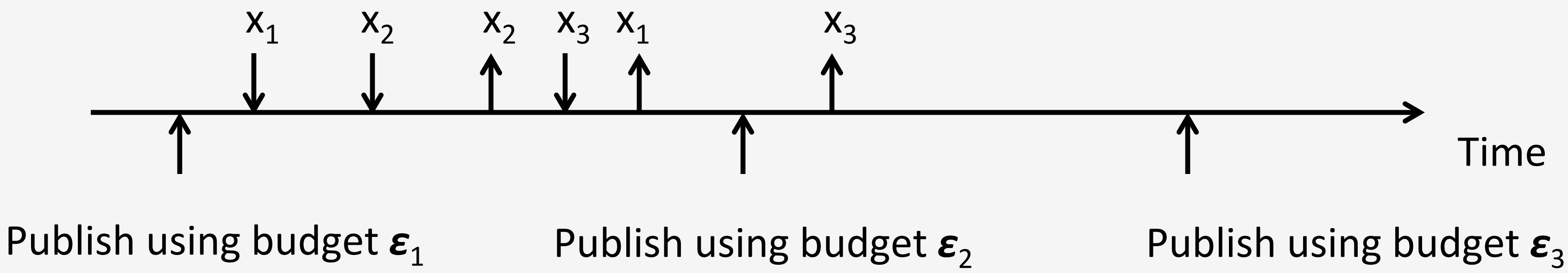




5. Dynamic Dataset

Scenario

- Over time, entities enter and leave the dataset D.


x_1
x_2
x_2
x_3
x_1
x_3
Time
Publish using budget $\boldsymbol{\varepsilon}_1$
Publish using budget $\boldsymbol{\varepsilon}_2$
Publish using budget $\boldsymbol{\varepsilon}_3$

- Suppose an entity must leave within a period of, say δ time. Then, both notion (standard and δ-neighbourhood) are equivalent.

- Similarly, if an entity stays for longer time, there is still assurance on the bound, but weaker.

- Under the notion of standard neighbourhood, continuous publishing with limited budgets would lead to lower utility.

- With the relaxed notion of δ-neighbourhood, we can now analyse the effect of continuous publishing. We can achieve *sustainable privacy.*

Offline vs Mechanism

- (offline) The data owner wants to publish at a few checkpoints over the time. The technical challenge is in distributing the budget over the checkpoints, so as to optimize the utility.

  We propose an algorithm for some choices of utility functions.

- (online) The data owner doesn't know when the data are to be published. When a request arrives, the data owner determines the budget and publishes the data immediately.

  We propose a online algorithm that choose the budget that optimize the average utility on a few sampled input.

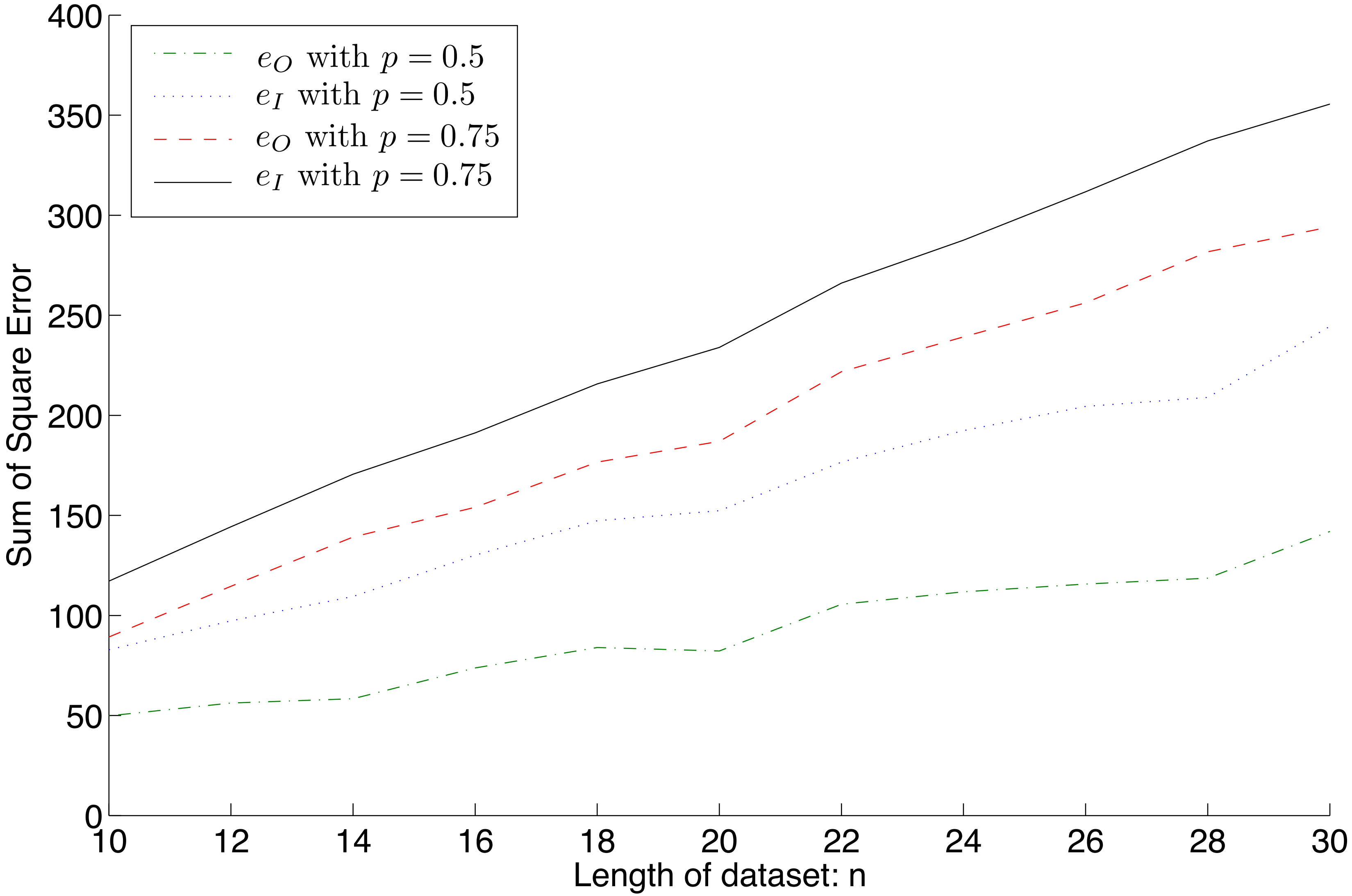
6. Conclusion

Conclusion

- We propose using the δ-neighbourhood. There are a few ways to interpret the relaxed notion.
  - When validity of datasets implies δ-neighbourhood, both notions are the same.
  - Redistribution of assurance with more stress on nearby entities.

- We propose a framework that exploits δ-neighbourhood for better utility.

- For dynamic datasets, we looked into the online vs offline settings.


e_O with $p = 0.5$
e_I with $p = 0.5$
e_O with $p = 0.75$
e_I with $p = 0.75$
Sum of Square Error
0
50
100
150
200
250
300
350
400
Length of dataset: n
10
12
14
16
18
20
22
24
26
28
30