VLDP - Geolife & Smart Meter dataset

Dataset preprocessing and computing parameters

In this notebook, we show how we created the datasets as used in the paper from the original open datasets. Moreover, we compute the value of the γ parameter for our use case evaluations.

Privacy Parameters

Before going into the specifics of the use cases, we first define the randomizer algorithms for real values and histograms (following the paper).

The code below also determines the privacy amplification we get through shuffling, for some arbitrarily chosen parameters. Give the target overall ϵ and δ and the number of participants n, it gives the ϵ_0 needed by the local randomizer.

We make use of the shuffleddp repository [Balle'19] to determine the bounds of our LDP randomizers and implement them in Python.

Let's first import all the required packages.

[Balle'19] Balle, B., Bell, J., Gascón, A. and Nissim, K., 2019. The privacy blanket of the shuffle model. In Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II 39 (pp. 638-667). Springer International Publishing.

In [15]:
```
import numpy as np
import csv
import os
import pandas as pd
import itertools
import kagglehub
from urllib.request import urlretrieve, urlcleanup
from geopy.geocoders import Nominatim
from geopy.exc import GeocoderTimedOut
from zipfile import ZipFile
import datetime
from shuffleddp.mechanisms import *
from shuffleddp.amplification_bounds import *
```

Algorithm for Real Values

Given $x \in [0, 1]$, the following algorithm calculates the sum of i such values by first encoding them with precision k and then applying the LDP and the Analyzer algorithm given in Section 4.1 of [Balle'19].

In [16]:
```
# Target (eps, delta)-guarantee required
eps = 0.1
```

```
delta = 1e-6

n = 5000  # number of participants
k = 100  # precision level
rrk = RRMechanism(k=k + 1)  # we have the range of {0, 1, ..., k + 1}

bound_types = [Hoeffding, BennettExact]
all_bounds = []
for B in bound_types:
    all_bounds.append(B(rrk))

print(f"Epsilon: {eps}", eps)
print(f"Delta: {delta}")
print(f"Number of participants: {n}")
bounds = {b.get_name(): b.get_eps0(eps, n, delta) for b in all_bounds}
print(f"Bounds: {bounds}")

gamma = rrk.get_gamma()[0]
print(f"Gamma: {gamma}")


# The first part of the randomizer (float encoding as int)
def encode(x, k):
    p = x * k - np.floor(x * k)
    x_enc = np.floor(x * k) + np.random.binomial(1, p)
    return x_enc


# second part of the randomizer (randomized response)
def RRMech(x, gamma, k):
    if not np.random.binomial(1, gamma):
        return x
    else:
        return np.random.randint(k + 1)


# apply float encoding to random inputs (as example)
true_vals = np.random.rand(1, n)
encode_v = np.vectorize(encode)
enc_true_vals = encode_v(true_vals, k)

# apply randomized response
RRMech_v = np.vectorize(RRMech)
received_vals = RRMech_v(enc_true_vals[0], gamma, k)

# compute outpus
sample_sum = sum(received_vals)
estimate = (sample_sum / k - gamma * n / 2) / (1 - gamma)
print(f"Estimate: {estimate}")
print(f"Actual: {sum(true_vals[0])}")
print(f"Received sum divided by k: {sample_sum / k}")
```

```
Epsilon: 0.1 0.1
Delta: 1e-06
Number of participants: 5000
Bounds: {'Hoeffding, RR-101': 0.9959842619145971, 'Bennett, RR-101': 2.552
3496569511277}
Gamma: 0.89509463046459
Estimate: 2524.307621347632
Actual: 2514.664336357696
Received sum divided by k: 2502.55
```

Algorithm for Histogram

Given $x \in [k]$, the following algorithm calculates the histogram of values by applying the LDP algorithm given in Section 3.1 of [Balle'19]. Note that the values are already integers.

In [17]:
```python
# Target (eps, delta)-guarantee required
eps = 0.2
delta = 1e-6

n = 1000  # number of participants
k = 100  # precision level
rrk = RRMechanism(k=k)  # we have the range of {0, 1, ..., k}

bound_types = [Hoeffding, BennettExact]
all_bounds = []
for B in bound_types:
    all_bounds.append(B(rrk))

print(f"Epsilon: {eps}", eps)
print(f"Delta: {delta}")
print(f"Number of participants: {n}")
bounds = {b.get_name(): b.get_eps0(eps, n, delta) for b in all_bounds}
print(f"Bounds: {bounds}")

gamma = rrk.get_gamma()[0]
print(f"Gamma: {gamma}")


# the randomizer algorithm (randomized response)
def RRMechHist(x, gamma, k):
    b = np.random.binomial(1, gamma)
    if not b:
        return x
    else:
        return np.random.randint(1, k + 1)


# generate random inputs
true_vals = np.random.choice(np.arange(1, k + 1), n)

# apply randomizer
RRMechHist_v = np.vectorize(RRMechHist)
received_vals = RRMechHist_v(true_vals, gamma, k)

# compute outputs
_unique, true_counts = np.unique(true_vals, return_counts=True)
_unique, est_counts = np.unique(received_vals, return_counts=True)
```

```
print(f"Estimate: {est_counts}")
print(f"Actual: {true_counts}")
```

```
Epsilon: 0.2 0.2
Delta: 1e-06
Number of participants: 1000
Bounds: {'Hoeffding, RR-100': 0.9026877389112389, 'Bennett, RR-100': 2.325
7610671469986}
Gamma: 0.9154619731616345
Estimate: [15 13 13 13 12  5 10 10 13  8  7 14  8 11  5 10  9 13  9  6  6
14  3 10
 11 12 11 13  8 10 18  7 17 16  8  8  9 10  8 15 11  8 14 13 10  5  9  7
 10 11  9  6  9  9  7 11 15  4  9  5 10 12  8  6  9 12  3 12 11 10  5 12
 11 11 12 12 11 11  8  5  7  8 11  9  6 14 13  9  2  9 17 12  9 19 11 15
 13  6 11  8]
Actual: [10 17 11 11 13 11 17  8  8 12  6  7 14  8  9 10 14 10  9  8 11 13
9 11
  8  6  9 10 10 14  9  6 13  6  7  8 10  9  8  6 10 12  9 14  8  6 10 16
  7  9  9  7  7 12 14  7 12  6 12  6  8 11 10  4 20 10 16 10 12  4  9 13
 10 19  7  4 10  8 11 12  8 18  8  7 13  8 15 12  8 15 11 10  5 13  8  8
  9  6 12  9]
```

Smart Meter Data (Use Case 1)

In this Section we describe the dataset preparation and show how we determine the bounds/run a DP example on this dataset.

The dataset is an application of the mechanism for summing up real numbers (from which we can obtain the average as well), i.e., Algorithms for Real Values, as mentioned above. The dataset is taken from: https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london. In particular, the dataset `daily_dataset.csv` is used.

Note: to download the original dataset from Kaggle using the code below, one MIGHT (often it is not needed) has to make an account and set up a token following the "Installation" and "Authentication" sections on https://www.kaggle.com/docs/api. One can then uncomment the specified line below and use the information from the token to login to the Kaggle API.

In [18]:
```
# NOTE: IF KAGGLE API KEY IS NEEDED UNCOMMENT BELOW
# kagglehub.login()

dataset_path = kagglehub.dataset_download("jeanmidev/smart-meters-in-lond
dataset = "daily_dataset.csv"
df = pd.read_csv(os.path.join(dataset_path, dataset))
```

In [19]:
```
# inspect dataset
df.head
```

Out[19]:
```
<bound method NDFrame.head of             LCLid         day  energy_median  \
ian  energy_mean  energy_max
0         MAC000131  2011-12-15         0.4850     0.432045       0.868
1         MAC000131  2011-12-16         0.1415     0.296167       1.116
2         MAC000131  2011-12-17         0.1015     0.189812       0.685
3         MAC000131  2011-12-18         0.1140     0.218979       0.676
4         MAC000131  2011-12-19         0.1910     0.325979       0.788
...             ...         ...            ...          ...         ...
3510428   MAC004977  2014-02-24         0.0950     0.118458       0.580
3510429   MAC004977  2014-02-25         0.0675     0.084208       0.176
3510430   MAC004977  2014-02-26         0.1080     0.120500       0.282
3510431   MAC004977  2014-02-27         0.0720     0.114062       0.431
3510432   MAC004977  2014-02-28         0.0970     0.097000       0.097

         energy_count  energy_std  energy_sum  energy_min
0                  22    0.239146       9.505       0.072
1                  48    0.281471      14.216       0.031
2                  48    0.188405       9.111       0.064
3                  48    0.202919      10.511       0.065
4                  48    0.259205      15.647       0.066
...               ...         ...         ...         ...
3510428            48    0.093814       5.686       0.052
3510429            48    0.037107       4.042       0.046
3510430            48    0.069332       5.784       0.046
3510431            48    0.094482       5.475       0.047
3510432             1         NaN       0.097       0.097

[3510433 rows x 9 columns]>
```

DP Parameters and Example Run (Use Case 1)

In [20]:
```
# maximum possible value of energy -- we will normalize using this
max_energy = df["energy_mean"].max()
print("Max energy:", max_energy)

# Total number of households
households = df["LCLid"].unique()
n = len(households)

eps = 0.2  # Target (eps, delta)-guarantee required
delta = 1e-6
k = 10  # precision level
rrk = RRMechanism(k=k + 1)

bound_types = [Hoeffding, BennettExact]
all_bounds = []
for B in bound_types:
    all_bounds.append(B(rrk))

print(f"Epsilon: {eps}", eps)
print(f"Delta: {delta}")
print(f"Number of participants: {n}")
bounds = {b.get_name(): b.get_eps0(eps, n, delta) for b in all_bounds}
print(f"Bounds: {bounds}")

gamma = rrk.get_gamma()[0]
print(f"Gamma: {gamma}")

num_days = int(1 / eps)  # we will run the mechanism a total of 1/eps tim
```

```
print("k:", k)
print("n:", n)
print("Number of days:", num_days)
print("eps0:", rrk.get_eps0())
print("================\n")

last_day = "2014-02-25"
cur_date = datetime.datetime.strptime(last_day, '%Y-%m-%d').date()
delta = datetime.timedelta(days=1)


# Normalize energy values within [0, 1]
def normalizeVals(vals):
    for i in range(len(vals)):
        vals[i] = vals[i] / (max_energy)
    return vals


# also store data for writing to csv
energy_vals_for_csv = []

# do an example DP run on this data and simultaneously parse the data
for i in range(num_days):
    day = cur_date.strftime('%Y-%m-%d')
    cur_date -= delta
    df0 = df[['LCLid', 'day', 'energy_mean']]
    df1 = df0[df0['day'] == day]
    energy_vals = [0.0 for i in range(len(households))]
    for j in range(len(households)):
        energy = df1.loc[df1['LCLid'] == households[j], 'energy_mean'].va
        if energy.size != 0:
            energy_vals[j] = energy[0]
    energy_vals = normalizeVals(energy_vals)
    true_vals = energy_vals
    encode_v = np.vectorize(encode)
    enc_true_vals = encode_v(true_vals, k)

    RRMech_v = np.vectorize(RRMech)
    received_vals = RRMech_v(enc_true_vals, gamma, k)

    sample_sum = sum(received_vals)
    # This is the de-biasing step in Algorithm 3 of [Balle'19]
    estimate = (sample_sum / k - gamma * n / 2) / (1 - gamma)
    print(f"Run {i + 1}:")
    print(f"Estimate: {estimate / n}")
    print(f"Actual: {sum(true_vals) / n}")
    print("==================\n")
    energy_vals_for_csv.append(energy_vals)
```

```
Max energy: 6.928250020833329
Epsilon: 0.2 0.2
Delta: 1e-06
Number of participants: 5566
Bounds: {'Hoeffding, RR-11': 1.6116776981605088, 'Bennett, RR-11': 2.48270
496095323}
Gamma: 0.5006005204469995
k: 10
n: 5566
Number of days: 5
eps0: 2.482704960952729
================

Run 1:
Estimate: 0.028539038174632547
Actual: 0.027832181900990727
==================

Run 2:
Estimate: 0.0018451019919219146
Actual: 0.028096308785484303
==================

Run 3:
Estimate: 0.020120734850165824
Actual: 0.031060984630618366
==================

Run 4:
Estimate: 0.030481723557201818
Actual: 0.029218913551385617
==================

Run 5:
Estimate: 0.013717068218733914
Actual: 0.028447346459788043
==================
```

Writing Extracted Smart Meter Data

The following extracts only relevant information into a CSV file. Namely the average energy consumption per household over the days used in the algorithm.

In [21]:
```
# Write the normalized energy values into a csv file
with open("energy_data.csv", "w", newline='') as f:
    wr = csv.writer(f, delimiter=",")
    header_row = ["household", "day", "average energy"]
    wr.writerow(header_row)
    for i in range(len(energy_vals_for_csv)):
        vals = energy_vals_for_csv[i]
        for j in range(len(vals)):
            row = [j, i, vals[j]]
            wr.writerow(row)
```

Geolife GPS Trajectory Dataset (Use Case 2)

In this Section we describe the dataset preparation and show how we determine the bounds/run a DP example on this dataset.

Taken from https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/. The following extracts the first longitude, latitude entry on a given date from files corresponding to all users. Not all users have date for each day.

First we download and unpack the original data. (Note: This can take up to 10-15 minutes, since it's a lot of data to unpack)

In [22]:
```
zip_path, _headers = urlretrieve(
    "https://download.microsoft.com/download/F/4/8/F4894AA5-FDBC-481E-928

with ZipFile(zip_path, 'r') as zip_file:
    zip_file.extractall(os.getcwd())

# remove downloaded tmp file
urlcleanup()
```

First Latitude, Longitude Reading from All Users One Day at a Time

We only need a subselection of the data, so we do that as follows. This code takes the first lat long from each user's file of the first 5 days.

In [23]:
```
work_dir = "Geolife Trajectories 1.3/Data"
users = ["{:03d}".format(i) for i in range(182)]
sub_dir = "Trajectory"

latLongs = []
days = 5

for user in users:
    cur_dir = os.path.join(work_dir, user, sub_dir)
    files = sorted([filename for filename in os.listdir(cur_dir)])
    for day in range(days):
        if day < len(files):
            cur_file = os.path.join(cur_dir, files[day])
            with open(cur_file, 'r') as f:
                reader = csv.reader(f, delimiter='|')
                rows = list(reader)

            flat_rows = itertools.chain.from_iterable(rows)
            list_rows = [i.strip().split(',') for i in flat_rows]

            df = pd.DataFrame(list_rows[6:])  # first 6 lines are useless

            lat = df.iloc[0][0]
            long = df.iloc[0][1]
            latLongs.append([user, day, lat, long])

print(f"Number of records: {len(latLongs)}")
print("Head:")
print(latLongs[:6])
```

```
Number of records: 851
Head:
[['000', 0, '39.984702', '116.318417'], ['000', 1, '40.008304', '116.31987
6'], ['000', 2, '39.907414', '116.370017'], ['000', 3, '39.994622', '116.3
26757'], ['000', 4, '40.01229', '116.297072'], ['001', 0, '39.984094', '11
6.319236']]
```

Calling the Reverse Geo API

Next we transform the latitudes and longitudes found in the data into postcodes using a geodata lookup.

NOTE: please specify the use agent on line 7, for example use your e-mail or the name of the application. This is necessary to follow the conditions of the API used.

In [24]:
```
# TODO: set USER AGENT
# initialize Nominatim API
geolocator = Nominatim(user_agent="PLEASE SPECIFY")

for idx, data in enumerate(latLongs):
    print(f"{idx + 1}/{len(latLongs)}")
    lat = data[2]
    long = data[3]
    try:
        location = geolocator.reverse(lat + "," + long, language='en')
        address = location.raw['address']
        if 'postcode' in address:
            data.append(address['postcode'])
        else:
            data.append("None")
    except GeocoderTimedOut as e:
        print("Error: geocode failed on input %s" % (lat + "," + long))
        data.append("TimeOut")
print("done")
```

```
1/851
2/851
3/851
4/851
5/851
6/851
7/851
8/851
9/851
10/851
11/851
12/851
13/851
14/851
15/851
16/851
17/851
18/851
19/851
20/851
21/851
22/851
23/851
24/851
25/851
26/851
27/851
28/851
29/851
30/851
31/851
32/851
33/851
34/851
35/851
36/851
37/851
38/851
39/851
40/851
41/851
42/851
43/851
44/851
45/851
46/851
47/851
48/851
49/851
50/851
51/851
52/851
53/851
54/851
55/851
56/851
57/851
58/851
59/851
60/851
```

```
61/851
62/851
63/851
64/851
65/851
66/851
67/851
68/851
69/851
70/851
71/851
72/851
73/851
74/851
75/851
76/851
77/851
78/851
79/851
80/851
81/851
82/851
83/851
84/851
85/851
86/851
87/851
88/851
89/851
90/851
91/851
92/851
93/851
94/851
95/851
96/851
97/851
98/851
99/851
100/851
101/851
102/851
103/851
104/851
105/851
106/851
107/851
108/851
109/851
110/851
111/851
112/851
113/851
114/851
115/851
116/851
117/851
118/851
119/851
120/851
```

121/851
122/851
123/851
124/851
125/851
126/851
127/851
128/851
129/851
130/851
131/851
132/851
133/851
134/851
135/851
136/851
137/851
138/851
139/851
140/851
141/851
142/851
143/851
144/851
145/851
146/851
147/851
148/851
149/851
150/851
151/851
152/851
153/851
154/851
155/851
156/851
157/851
158/851
159/851
160/851
161/851
162/851
163/851
164/851
165/851
166/851
167/851
168/851
169/851
170/851
171/851
172/851
173/851
174/851
175/851
176/851
177/851
178/851
179/851
180/851

```
181/851
182/851
183/851
184/851
185/851
186/851
187/851
188/851
189/851
190/851
191/851
192/851
193/851
194/851
195/851
196/851
197/851
198/851
199/851
200/851
201/851
202/851
203/851
204/851
205/851
206/851
207/851
208/851
209/851
210/851
211/851
212/851
213/851
214/851
215/851
216/851
217/851
218/851
219/851
220/851
221/851
222/851
223/851
224/851
225/851
226/851
227/851
228/851
229/851
230/851
231/851
232/851
233/851
234/851
235/851
236/851
237/851
238/851
239/851
240/851
```

```
241/851
242/851
243/851
244/851
245/851
246/851
247/851
248/851
249/851
250/851
251/851
252/851
253/851
254/851
255/851
256/851
257/851
258/851
259/851
260/851
261/851
262/851
263/851
264/851
265/851
266/851
267/851
268/851
269/851
270/851
271/851
272/851
273/851
274/851
275/851
276/851
277/851
278/851
279/851
280/851
281/851
282/851
283/851
284/851
285/851
286/851
287/851
288/851
289/851
290/851
291/851
292/851
293/851
294/851
295/851
296/851
297/851
298/851
299/851
300/851
```

```
301/851
302/851
303/851
304/851
305/851
306/851
307/851
308/851
309/851
310/851
311/851
312/851
313/851
314/851
315/851
316/851
317/851
318/851
319/851
320/851
321/851
322/851
323/851
324/851
325/851
326/851
327/851
328/851
329/851
330/851
331/851
332/851
333/851
334/851
335/851
336/851
337/851
338/851
339/851
340/851
341/851
342/851
343/851
344/851
345/851
346/851
347/851
348/851
349/851
350/851
351/851
352/851
353/851
354/851
355/851
356/851
357/851
358/851
359/851
360/851
```

361/851
362/851
363/851
364/851
365/851
366/851
367/851
368/851
369/851
370/851
371/851
372/851
373/851
374/851
375/851
376/851
377/851
378/851
379/851
380/851
381/851
382/851
383/851
384/851
385/851
386/851
387/851
388/851
389/851
390/851
391/851
392/851
393/851
394/851
395/851
396/851
397/851
398/851
399/851
400/851
401/851
402/851
403/851
404/851
405/851
406/851
407/851
408/851
409/851
410/851
411/851
412/851
413/851
414/851
415/851
416/851
417/851
418/851
419/851
420/851

421/851
422/851
423/851
424/851
425/851
426/851
427/851
428/851
429/851
430/851
431/851
432/851
433/851
434/851
435/851
436/851
437/851
438/851
439/851
440/851
441/851
442/851
443/851
444/851
445/851
446/851
447/851
448/851
449/851
450/851
451/851
452/851
453/851
454/851
455/851
456/851
457/851
458/851
459/851
460/851
461/851
462/851
463/851
464/851
465/851
466/851
467/851
468/851
469/851
470/851
471/851
472/851
473/851
474/851
475/851
476/851
477/851
478/851
479/851
480/851

481/851
482/851
483/851
484/851
485/851
486/851
487/851
488/851
489/851
490/851
491/851
492/851
493/851
494/851
495/851
496/851
497/851
498/851
499/851
500/851
501/851
502/851
503/851
504/851
505/851
506/851
507/851
508/851
509/851
510/851
511/851
512/851
513/851
514/851
515/851
516/851
517/851
518/851
519/851
520/851
521/851
522/851
523/851
524/851
525/851
526/851
527/851
528/851
529/851
530/851
531/851
532/851
533/851
534/851
535/851
536/851
537/851
538/851
539/851
540/851

```
541/851
542/851
543/851
544/851
545/851
546/851
547/851
548/851
549/851
550/851
551/851
552/851
553/851
554/851
555/851
556/851
557/851
558/851
559/851
560/851
561/851
562/851
563/851
564/851
565/851
566/851
567/851
568/851
569/851
570/851
571/851
572/851
573/851
574/851
575/851
576/851
577/851
578/851
579/851
580/851
581/851
582/851
583/851
584/851
585/851
586/851
587/851
588/851
589/851
590/851
591/851
592/851
593/851
594/851
595/851
596/851
597/851
598/851
599/851
600/851
```

601/851
602/851
603/851
604/851
605/851
606/851
607/851
608/851
609/851
610/851
611/851
612/851
613/851
614/851
615/851
616/851
617/851
618/851
619/851
620/851
621/851
622/851
623/851
624/851
625/851
626/851
627/851
628/851
629/851
630/851
631/851
632/851
633/851
634/851
635/851
636/851
637/851
638/851
639/851
640/851
641/851
642/851
643/851
644/851
645/851
646/851
647/851
648/851
649/851
650/851
651/851
652/851
653/851
654/851
655/851
656/851
657/851
658/851
659/851
660/851

661/851
662/851
663/851
664/851
665/851
666/851
667/851
668/851
669/851
670/851
671/851
672/851
673/851
674/851
675/851
676/851
677/851
678/851
679/851
680/851
681/851
682/851
683/851
684/851
685/851
686/851
687/851
688/851
689/851
690/851
691/851
692/851
693/851
694/851
695/851
696/851
697/851
698/851
699/851
700/851
701/851
702/851
703/851
704/851
705/851
706/851
707/851
708/851
709/851
710/851
711/851
712/851
713/851
714/851
715/851
716/851
717/851
718/851
719/851
720/851

721/851
722/851
723/851
724/851
725/851
726/851
727/851
728/851
729/851
730/851
731/851
732/851
733/851
734/851
735/851
736/851
737/851
738/851
739/851
740/851
741/851
742/851
743/851
744/851
745/851
746/851
747/851
748/851
749/851
750/851
751/851
752/851
753/851
754/851
755/851
756/851
757/851
758/851
759/851
760/851
761/851
762/851
763/851
764/851
765/851
766/851
767/851
768/851
769/851
770/851
771/851
772/851
773/851
774/851
775/851
776/851
777/851
778/851
779/851
780/851

781/851
782/851
783/851
784/851
785/851
786/851
787/851
788/851
789/851
790/851
791/851
792/851
793/851
794/851
795/851
796/851
797/851

```
---------------------------------------------------------------------------
TimeoutError                              Traceback (most recent call last)
File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connectionpool.py:534, in HTTPConnectionPool._make_request(self, conn, method, url, body, headers, retries, timeout, chunked, response_conn, preload_content, decode_content, enforce_content_length)
    533 try:
--> 534     response = conn.getresponse()
    535 except (BaseSSLError, OSError) as e:

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connection.py:516, in HTTPConnection.getresponse(self)
    515 # Get the response from http.client.HTTPConnection
--> 516 httplib_response = super().getresponse()
    518 try:

File ~/.pyenv/versions/3.11.8/lib/python3.11/http/client.py:1390, in HTTPConnection.getresponse(self)
   1389 try:
-> 1390     response.begin()
   1391 except ConnectionError:

File ~/.pyenv/versions/3.11.8/lib/python3.11/http/client.py:325, in HTTPResponse.begin(self)
    324 while True:
--> 325     version, status, reason = self._read_status()
    326     if status != CONTINUE:

File ~/.pyenv/versions/3.11.8/lib/python3.11/http/client.py:286, in HTTPResponse._read_status(self)
    285 def _read_status(self):
--> 286     line = str(self.fp.readline(_MAXLINE + 1), "iso-8859-1")
    287     if len(line) > _MAXLINE:

File ~/.pyenv/versions/3.11.8/lib/python3.11/socket.py:706, in SocketIO.readinto(self, b)
    705 try:
--> 706     return self._sock.recv_into(b)
    707 except timeout:

File ~/.pyenv/versions/3.11.8/lib/python3.11/ssl.py:1314, in SSLSocket.recv_into(self, buffer, nbytes, flags)
   1311         raise ValueError(
   1312           "non-zero flags not allowed in calls to recv_into() on %s" %
   1313           self.__class__)
-> 1314     return self.read(nbytes, buffer)
   1315 else:

File ~/.pyenv/versions/3.11.8/lib/python3.11/ssl.py:1166, in SSLSocket.read(self, len, buffer)
   1165 if buffer is not None:
-> 1166     return self._sslobj.read(len, buffer)
   1167 else:

TimeoutError: The read operation timed out

The above exception was the direct cause of the following exception:
```

```
ReadTimeoutError                          Traceback (most recent call las
t)
File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connect
ionpool.py:787, in HTTPConnectionPool.urlopen(self, method, url, body, hea
ders, retries, redirect, assert_same_host, timeout, pool_timeout, release_
conn, chunked, body_pos, preload_content, decode_content, **response_kw)
    786 # Make the request on the HTTPConnection object
--> 787 response = self._make_request(
    788     conn,
    789     method,
    790     url,
    791     timeout=timeout_obj,
    792     body=body,
    793     headers=headers,
    794     chunked=chunked,
    795     retries=retries,
    796     response_conn=response_conn,
    797     preload_content=preload_content,
    798     decode_content=decode_content,
    799     **response_kw,
    800 )
    802 # Everything went great!

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connect
ionpool.py:536, in HTTPConnectionPool._make_request(self, conn, method, ur
l, body, headers, retries, timeout, chunked, response_conn, preload_conten
t, decode_content, enforce_content_length)
    535 except (BaseSSLError, OSError) as e:
--> 536     self._raise_timeout(err=e, url=url, timeout_value=read_timeou
t)
    537     raise

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connect
ionpool.py:367, in HTTPConnectionPool._raise_timeout(self, err, url, timeo
ut_value)
    366 if isinstance(err, SocketTimeout):
--> 367     raise ReadTimeoutError(
    368         self, url, f"Read timed out. (read timeout={timeout_valu
e})"
    369     ) from err
    371 # See the above comment about EAGAIN in Python 3.

ReadTimeoutError: HTTPSConnectionPool(host='nominatim.openstreetmap.org',
port=443): Read timed out. (read timeout=1)

The above exception was the direct cause of the following exception:

MaxRetryError                             Traceback (most recent call las
t)
File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/requests/adapte
rs.py:667, in HTTPAdapter.send(self, request, stream, timeout, verify, cer
t, proxies)
    666 try:
--> 667     resp = conn.urlopen(
    668         method=request.method,
    669         url=url,
    670         body=request.body,
    671         headers=request.headers,
    672         redirect=False,
```

```
    673             assert_same_host=False,
    674             preload_content=False,
    675             decode_content=False,
    676             retries=self.max_retries,
    677             timeout=timeout,
    678             chunked=chunked,
    679         )
    681 except (ProtocolError, OSError) as err:

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connect
ionpool.py:871, in HTTPConnectionPool.urlopen(self, method, url, body, hea
ders, retries, redirect, assert_same_host, timeout, pool_timeout, release_
conn, chunked, body_pos, preload_content, decode_content, **response_kw)
    868     log.warning(
    869         "Retrying (%r) after connection broken by '%r': %s", retri
es, err, url
    870     )
--> 871     return self.urlopen(
    872         method,
    873         url,
    874         body,
    875         headers,
    876         retries,
    877         redirect,
    878         assert_same_host,
    879         timeout=timeout,
    880         pool_timeout=pool_timeout,
    881         release_conn=release_conn,
    882         chunked=chunked,
    883         body_pos=body_pos,
    884         preload_content=preload_content,
    885         decode_content=decode_content,
    886         **response_kw,
    887     )
    889 # Handle redirect?

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connect
ionpool.py:871, in HTTPConnectionPool.urlopen(self, method, url, body, hea
ders, retries, redirect, assert_same_host, timeout, pool_timeout, release_
conn, chunked, body_pos, preload_content, decode_content, **response_kw)
    868     log.warning(
    869         "Retrying (%r) after connection broken by '%r': %s", retri
es, err, url
    870     )
--> 871     return self.urlopen(
    872         method,
    873         url,
    874         body,
    875         headers,
    876         retries,
    877         redirect,
    878         assert_same_host,
    879         timeout=timeout,
    880         pool_timeout=pool_timeout,
    881         release_conn=release_conn,
    882         chunked=chunked,
    883         body_pos=body_pos,
    884         preload_content=preload_content,
    885         decode_content=decode_content,
    886         **response_kw,
```

```
    887         )
    889 # Handle redirect?

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/connectionpool.py:841, in HTTPConnectionPool.urlopen(self, method, url, body, headers, retries, redirect, assert_same_host, timeout, pool_timeout, release_conn, chunked, body_pos, preload_content, decode_content, **response_kw)
    839     new_e = ProtocolError("Connection aborted.", new_e)
--> 841 retries = retries.increment(
    842     method, url, error=new_e, _pool=self, _stacktrace=sys.exc_info()[2]
    843 )
    844 retries.sleep()

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/urllib3/util/retry.py:519, in Retry.increment(self, method, url, response, error, _pool, _stacktrace)
    518     reason = error or ResponseError(cause)
--> 519     raise MaxRetryError(_pool, url, reason) from reason  # type: ignore[arg-type]
    521 log.debug("Incremented Retry for (url='%s'): %r", url, new_retry)

MaxRetryError: HTTPSConnectionPool(host='nominatim.openstreetmap.org', port=443): Max retries exceeded with url: /reverse?lat=39.975061&lon=116.329201&format=json&accept-language=en&addressdetails=1 (Caused by ReadTimeoutError("HTTPSConnectionPool(host='nominatim.openstreetmap.org', port=443): Read timed out. (read timeout=1)"))

During handling of the above exception, another exception occurred:

ConnectionError                           Traceback (most recent call last)
File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/geopy/adapters.py:482, in RequestsAdapter._request(self, url, timeout, headers)
    481 try:
--> 482     resp = self.session.get(url, timeout=timeout, headers=headers)
    483 except Exception as error:

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/requests/sessions.py:602, in Session.get(self, url, **kwargs)
    601 kwargs.setdefault("allow_redirects", True)
--> 602 return self.request(      , url, **kwargs)

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/requests/sessions.py:589, in Session.request(self, method, url, params, data, headers, cookies, files, auth, timeout, allow_redirects, proxies, hooks, stream, verify, cert, json)
    588 send_kwargs.update(settings)
--> 589 resp = self.send(prep, **send_kwargs)
    591 return resp

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/requests/sessions.py:703, in Session.send(self, request, **kwargs)
    702 # Send the request
--> 703 r = adapter.send(request, **kwargs)
    705 # Total elapsed time of the request (approximately)

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/requests/adapters.py:700, in HTTPAdapter.send(self, request, stream, timeout, verify, cert, proxies)
```

```
    698         raise SSLError(e, request=request)
--> 700     raise ConnectionError(e, request=request)
    702 except ClosedPoolError as e:

ConnectionError: HTTPSConnectionPool(host='nominatim.openstreetmap.org', p
ort=443): Max retries exceeded with url: /reverse?lat=39.975061&lon=116.32
9201&format=json&accept-language=en&addressdetails=1 (Caused by ReadTimeou
tError("HTTPSConnectionPool(host='nominatim.openstreetmap.org', port=443):
Read timed out. (read timeout=1)"))

During handling of the above exception, another exception occurred:

GeocoderUnavailable                       Traceback (most recent call las
t)
Cell In[24], line 10
      8 long = data[3]
      9 try:
---> 10     location = geolocator.reverse(lat +     + long, language=    )
     11     address = location.raw['address']
     12     if 'postcode' in address:

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/geopy/geocoder
s/nominatim.py:372, in Nominatim.reverse(self, query, exactly_one, timeou
t, language, addressdetails, zoom, namedetails)
    370 logger.debug("%s.reverse: %s", self.__class__.__name__, url)
    371 callback = partial(self._parse_json, exactly_one=exactly_one)
--> 372 return self._call_geocoder(url, callback, timeout=timeout)

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/geopy/geocoder
s/base.py:368, in Geocoder._call_geocoder(self, url, callback, timeout, is
_json, headers)
    366 try:
    367     if is_json:
--> 368         result = self.adapter.get_json(url, timeout=timeout, heade
rs=req_headers)
    369     else:
    370         result = self.adapter.get_text(url, timeout=timeout, heade
rs=req_headers)

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/geopy/adapters.
py:472, in RequestsAdapter.get_json(self, url, timeout, headers)
    471 def get_json(self, url, *, timeout, headers):
--> 472     resp = self._request(url, timeout=timeout, headers=headers)
    473     try:
    474         return resp.json()

File ~/src/reserch/VLDP/.venv/lib/python3.11/site-packages/geopy/adapters.
py:494, in RequestsAdapter._request(self, url, timeout, headers)
    492         raise GeocoderServiceError(message)
    493     else:
--> 494         raise GeocoderUnavailable(message)
    495 elif isinstance(error, requests.Timeout):
    496     raise GeocoderTimedOut("Service timed out")

GeocoderUnavailable: HTTPSConnectionPool(host='nominatim.openstreetmap.or
g', port=443): Max retries exceeded with url: /reverse?lat=39.975061&lon=1
16.329201&format=json&accept-language=en&addressdetails=1 (Caused by ReadT
imeoutError("HTTPSConnectionPool(host='nominatim.openstreetmap.org', port=
443): Read timed out. (read timeout=1)"))
```

In [ ]:
```
print("Head:")
print(latLongs[:6])
```

Data Preparation for CSV

In the code below we make our data ready for the use case. First, we condense the list of postcodes to the top 7 most used ones. The other are aggregated under "all_others". Then we will the missing entries with the "all_others" category as well and write the resulting data to a CSV file.

Note: the resulting .csv file might be slightly different from the one we created, the geodata api calls are not always consistent (timeouts may happen), so this could cause some changes. The overall file will be very similar though.

In [ ]:
```
df = pd.DataFrame(latLongs, columns=["User", "day", "lat", "long", "postc

# reduce the number of postcodes to the top 7, replace rest by "all_other
top_postcodes = df.value_counts(subset="postcode").drop("None").nlargest(
df["postcode"] = df["postcode"].mask(df["postcode"].isin(top_postcodes) =

# fill missing data
users = ["{:03d}".format(i) for i in range(182)]
days = 5
for day in range(days):
    for user in users:
        if not any((df["User"] == user) & (df["day"] == day)):
            df.loc[len(df)] = [user, day, None, None, "all_others"]

# Write to CSV
df.to_csv("geolife-postcodes-condensed-empties.csv", index=False)
```

Inspect the final dataset distribution

Below we look at the true distribution of the resulting dataset

In [ ]:
```
df = pd.read_csv("geolife-postcodes-condensed-empties.csv")
days = df['day'].unique()
users = df['User'].unique()
counts = {}
for day in days:
    counts[day] = {}
    postcodes = df.loc[df['day'] == day, 'postcode'].unique()
    for postcode in postcodes:
        counts[day][postcode] = df[(df['day'] == day) & (df['postcode'] =

for day in counts:
    c = 0
    for k, v in counts[day].items():
        print(k, v)
        c = c + v
    print(f"This count: {c}")
    print()
```

DP Parameters and Example Run (Use Case 2)

In [ ]:
```
# Total number of users
users = df["User"].unique()
n = len(users)
postcodes = df["postcode"].unique()
k = len(postcodes)  # histogram of postcodes
eps = 2  # Target (eps, delta)-guarantee required
# over multiple runs the eps add up, e.g., 5 runs => 5*eps
delta = 1e-4

rrk = RRMechanism(k=k)  # we have the range of {0, 1, ..., k}

bound_types = [Hoeffding, BennettExact]
all_bounds = []
for B in bound_types:
    all_bounds.append(B(rrk))

print(f"Epsilon: {eps}", eps)
print(f"Delta: {delta}")
print(f"Number of participants: {n}")
bounds = {b.get_name(): b.get_eps0(eps, n, delta) for b in all_bounds}
print(f"Bounds: {bounds}")

gamma = rrk.get_gamma()[0]
print(f"Gamma: {gamma}")

days = df["day"].unique()


def RRMech(x, gamma, postcodes):
    if not np.random.binomial(1, gamma):
        return x
    else:
        return np.random.choice(postcodes)


orig_dict = {}
syn_dict = {}

# do an example run
i = 0
for day in days:
    i += 1
    orig_dict[day] = {}
    syn_dict[day] = {}
    df0 = df[['User', 'day', 'postcode']]
    df1 = df0[df0['day'] == day]
    for user in users:
        postcode = df1.loc[df1['User'] == user, 'postcode'].values
        #print(postcode)
        if postcode.size != 0:
            postcode = postcode[0]
            if postcode in orig_dict[day]:
                orig_dict[day][postcode] += 1
            else:
                orig_dict[day][postcode] = 1
            priv_postcode = RRMech(postcode, gamma, postcodes)
            if priv_postcode in syn_dict[day]:
                syn_dict[day][priv_postcode] += 1
            else:
```

```
                syn_dict[day][priv_postcode] = 1

    print("Run " + str(i) + ":")
    print("Postcode, DP, Original:")
    print("==========\n")
    for k, v1 in syn_dict[day].items():
        v2 = orig_dict[day][k]
        print(k + ", " + str(v1) + ", " + str(v2))
```

ノートブック全体の構成と“何をしているか”

この `LDP-Shuffle-Parameters.ipynb` は、**ローカル差分プライバシ (LDP) → シャッフルモデル** に移行したときに

- **どの乱数化パラメータ (ε₀, γ)** をクライアント側で使えば
- **全体として所望の (ε, δ) プライバシ保証** が得られるか を実データ上で決定し、サンプル実行で推定精度を確かめる「再現用スクリプト」です。 以下では **セル単位** に具体的な処理と数式・理由を掘り下げます。

---

1. 共通の前提

記号	意味	ノートブックでの採用値
n	同時参加クライアント数	GeoLife = 182, Smart-Meter ≈ 5,566
k	量子化精度（実数）またはカテゴリ数（ヒストグラム）	実数 k = 10 or 100、ヒスト k = 8
ε, δ	システム全体（シャッフル後）のプライバシ保証	GeoLife ε = 2, δ = 1e-4 Smart-Meter ε = 0.2, δ = 1e-6
ε₀	各クライアントが使う LDP 乱数化器の強さ	Hoeffding/Bennett 境界で逆算
γ	「真値を送る確率」= (e^{ε₀}) / (e^{ε₀}+k-1)	ライブラリが自動計算

なぜ ε₀ を強くしても良いのか？

シャッフルサーバが **発信元匿名化** することで、複数クライアントの出力が混ざり プライバシが “増幅” される (Balle et al., CRYPTO 2019)。 したがって「ターゲット ε」に収まるまで ε₀ を（ある程度）緩めて計算量を下げられます。

---

2. 実数値サマリー用アルゴリズム（Smart-Meter）

2-1. 値の量子化 `encode(x, k)`

$$x \in [0,1] \xrightarrow{\text{確率的丸め}} \tilde{x} \in \{0,1,\dots,k\}$$

```
p       = x*k - floor(x*k)          # 余り
x_enc   = floor(x*k) + Bernoulli(p)
```

整数化することで **k+1** 通りのカテゴリに落とし込み、後段 Randomized Response を適用できる。

2-2. k-ary Randomized Response `RRMech`

$$\hat{x} = \begin{cases} \tilde{x} & \text{確率 } \gamma \\ \mathrm{Unif}\{0,\dots,k\} & \text{確率 } 1-\gamma \end{cases}$$

ここで

$$\gamma = \frac{e^{\varepsilon_0}}{e^{\varepsilon_0} + k}$$

2-3. 集計側のデバイアス推定

受信値の総和 $s = \sum_i \hat{x}_i$ から真の総和 $\sum x_i$ を推定：

$$\widehat{\mathrm{SUM}} = \frac{s/k - \gamma n/2}{1-\gamma}$$

理由：乱数化器の期待値 $E[\hat{x}] = \gamma\, E[\tilde{x}] + (1-\gamma)\, k/2$ を逆解消。

2-4. 増幅境界で ε₀ を決定

ライブラリ `shuffleddp.amplification_bounds` が

$$\mathrm{eps0} = B^{-1}(\varepsilon, n, \delta)$$

を Hoeffding 版／Bennett 精密版で数値的に返す。 今回 Bennett → ε₀≈2.48 なので γ≈0.50。

2-5. ノートブックでの検証

- **5 日分 (1/ε) をループ**
- 各日 about 5,500 値を乱数化 → 集計 → 上式でデバイアス
- **真値平均との誤差を出力**（典型的に 1e-3〜1e-2）

---

3. ヒストグラム用アルゴリズム（GeoLife）

緯度経度 → 郵便番号へ逆ジオコーディングし、**出現上位7 + all_others** の 計 **k = 8** 区分に圧縮。

3-1. Randomized Response for categorical

$$\hat{x} = \begin{cases} x & \text{確率 } \gamma \\ \mathrm{Unif}(k) & \text{確率 } 1-\gamma \end{cases}$$

推定: 各カテゴリ c について

$$\widehat{N_c} = \frac{n_c - (1-\gamma)\, n/k}{\gamma}$$

（n_c: 受信カウント）

3-2. 実行フロー

1. **182 ユーザ × 5 日** の (user, day) 行を埋めて欠損を all_others に。
2. 目標 ε = 2 なので ε₀≈0.90, γ≈0.91 とかなり「真値温存」寄り。
3. 1 日ごとに乱数化 → 集計→推定を印字。

---

4. データ前処理セクション

セクション	主な I/O	注意点
Smart-Meter 1.0.5–1.0.7	Kaggle CSV → `energy_data.csv`	`energy_mean` を最大値で正規化し 0–1 射影。
GeoLife 1.0.8–1.0.12	ZIP → 逆ジオ→ `geolife-postcodes-condensed-empties.csv`	`geopy` API がタイムアウトすると `postcode=TimeOut/None` 。最終的にカテゴリ 8 件に集約。

---

5. 例外処理・実装上のハマりどころ

- **逆ジオコーディング** は 1 秒未満のタイムアウト指定なので連続失敗時に `GeocoderUnavailable` 。　→ 実際の実験では Postcode キャッシュ or 有料 API を推奨。
- **ε₀ > 1** になると **γ が 0.5 近辺まで落ち**、通信量は増えないが推定分散が拡大。
  Smart-Meter は n が大きいので許容、GeoLife は k 小さい＆ε 高いので γ↑。
- **量子化誤差:** k が小さいと高分解能情報が失われる。Smart-Meter で k = 10 なのは「電力平均を1桁分解能で十分」と判断した設計パラメータ。

まとめ ― このノートで“出来ること”

1. **プライバシ増幅境界を使った ε₀, γ の自動算出**
2. **実データに LDP → シャッフルモデルを適用**し、推定精度を即座に確認
3. **学習用 CSV を生成** ― 論文の実験コードはこの出力をそのまま読む

この一冊で「VLDP のパラメータ選定〜簡易再現実験」が完結する設計になっています。