Carnegie Mellon University
Department of Computer Science
15-826 Multimedia and Data Mining
C. Faloutsos, Fall 2025

# Homework 2

Due: on canvas, at **2:00pm, on Fri 09/19/2025**

**VERY IMPORTANT:**

- Deposit **pdf copy** of your answers, on canvas.

**Reminders:**

- *Plagiarism*: Homework is to be completed *individually*.
- *Typeset* your answers. You may use the pdf of the handout, to type/circle your answers. Illegible handwriting may get zero points.
- *Late homeworks*: Please follow the instructions here

**For your information:**

- Graded out of **100** points; **4** questions total
- Rough time estimate: *2-6 hours*
- Weight: 3% of total course grade.

*Revision* : 2025/09/12 01:41

| Question | Points | Score |
|---|---|---|
| Warm-up: duckdb | 60 | |
| Z-/Hilbert ordering | 15 | |
| R-trees | 15 | |
| Graph patterns | 10 | |
| Total: | 100 | |

## Question 1: Warm-up: duckdb ........................ [60 points]

The goal is to become familiar with the properties of duckdb: it is able to answer sql queries on compressed csv files, and even do joins on two of them.

**Reminder:** The phonecall data etc are all in the box directory We shall refer to it as **DATA_HOME** in this question.

**Data** As mentioned in the canvas discussion, you are given a script that anonymizes the long sha_hashcodes, into 1-$n$ integers.

- *original file*: For this homework we will use the data of the first two days of November: DATA_HOME/01_data/CMU_2021-11-01_02_cmu.csv.gz We shall refer to it as the *original* file.
- *trimmed file*: For your convenience, we run the anonymizer script and put the result into file DATA_HOME/02_subsets/CMU_2021-11-01_02_trimmed.csv.gz has such short integers for callers and call-examples (and we also omited a few not-so-useful columns).
- *lookup table*: The anonymizer script also produced a look-up table DATA_HOME/02_subsets/CMU_2021-11-01_02_lookup.csv.gz with the obvious two columns: original sha256_code, and our integer id).

**Sample queries:** Also for your convenience, we provide some queries at DATA_HOME/04_auxiliary/duckdb_for_hw2, with a 'makefile' and self-explanatory names.

**Your tasks:**

(a) [**10 points**] Run the query *04_outCalls.sql*, and report the integer-ids and number of phonecalls for the top 10 heavy-hitter, that is, the ones with the most out-going phonecalls.

(b) [**10 points**] Run the query *11_revealHeavyHitters.sql*: That is, as above, but reveal their sha256 codes (instead of our integer-ids). Report the results.

(c) [**2 points**] What was the wall-clock time of the above query?

(d) [**10 points**] Run the query *12_HeavyHitters_from_original.sql*. Report the results.

(e) [**3 points**] Report the wall-clock time of the above query. Was it slower than the query on the *trimmed* file?

(f) [**10 points**] Write a query to find the top 10 most talkative callers - report their sha256 code, and their total duration of their (out-going) phonecalls. Again, sort by duration descending; break ties by sha256 code ascending.

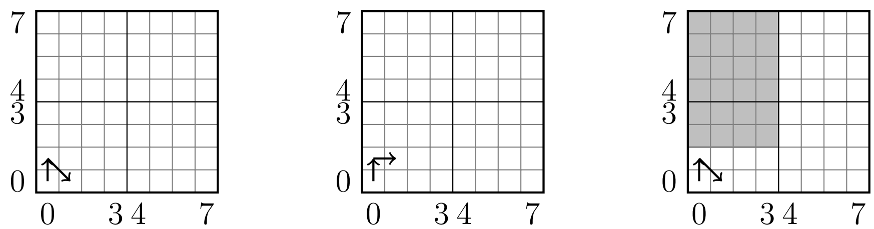
(g) [**15 points**] Give the SQL (duckdb) code of your 'talkative callers' query.

## Question 2: Z-/Hilbert ordering .................... [15 points]

Consider a $2^n \times 2^n$ grid, and the z-curve on it. As usually, its first step is *vertical*, that is:

- the (0,0) cell has decimal z-value = 0
- the (0,1) cell is next, with decimal z-value = 1

Figure 1(a) shows the first two steps (arrow) of such a z-curve, on an 8×8 grid (which obviously has ranges (0,7) × (0,7) ).




(a) z-values  (b) hilbert-values  (c) z-values for shaded area

Figure 1: Grids, for z- and hilbert-values

(a) [**2 points**] Which is the cell with the highest z-value? (Give (x,y) coordinates, like, e.g., (3,3))

(a) ______________

(b) [**5 points**] Which is the cell with the highest **hilbert**-value? (Figure 1(b) shows the first two steps of the curve).

(b) ______________

(c) [**3 points**] How many z-values do we need for the shaded area of Figure 1(c)? (Recall that '*'s can only be at the end - that is, e.g., 01**01 is not valid.)

(c) ______________

(d) [**5 points**] Give the z-value(s) for the shaded area of Figure 1(c)

........................................................................................................
........................................................................................................
........................................................................................................
........................................................................................................
........................................................................................................
........................................................................................................
........................................................................................................
........................................................................................................

## Question 3: R-trees ................................ [15 points]

The foils, and Pagel's formula, refers to intersection queries. Here, we focus on *inclusion* queries: A query $Q$ (shaded rectangle in Figure 2) would retrieve all the rectangles that completely contain it, like rectangle $A$.

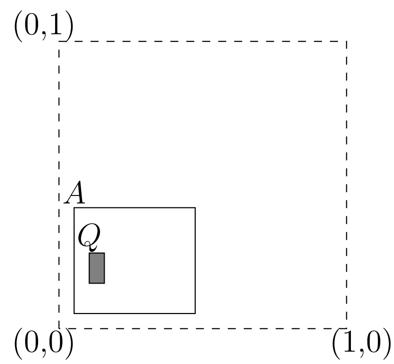



Figure 2: Example of inclusion query, in 2-d: the white rectangle $A$ includes the shaded query rectangle $Q$, and thus it qualifies under the inclusion query with $Q$.

Consider a **3-d** setting. Consider a parent node $P$ in an R-tree, whose MBR has sides $x_1$=0.2, $x_2$=0.3 and $x_3$=0.4. Consider also a query $Q$ with sides $(q_1, q_2, q_3)$. Assume the following:

- Everything is in the unit **cube**
- Assume that $x_i > q_i$ $(i = 1, 2, 3)$
- As in the foils, the query center is uniformly distributed in the available space, with wrapping around (as Pagel's formula also assumes).

(a) [**2 points**] What is the probability that parent $P$ will be retrieved, under a point query $(q_1 = q_2 = q_3 = 0)$

(a) ____________

(b) [**5 points**] What is the formula for the probability that a parent node with dimensions $x_1, x_2, x_3$, will completely contain the query box $Q$ with dimensions $q_1, q_2, q_3$.

A. $(x_1 + q_1) * (x_2 + q_2) * (x_3 + q_3)$
B. $(x_1/q_1) * (x_2/q_2) * (x_3/q_3)$
C. $(q_1/x_1) * (q_2/x_2) * (q_3/x_3)$
D. $(x_1 - q_1) * (x_2 - q_2) * (x_3 - q_3)$
E. $(q_1 - x_1) * (q_2 - x_2) * (q_3 - x_3)$
F. None of the above - the correct formula is: .........

(c) [**8 points**] Compute the probability that the query $Q$ with sides $q_1 = q_2 = q_3 = 0.1$, will be completely inside parent node $P$ ($x_1$=0.2, $x_2$=0.3 and $x_3$=0.4)
Give the exact, *numerical* answer.

(c) ____________

## Question 4: Graph patterns ......................... [10 points]

(a) **[4 points]** Suppose you are monitoring a million-scale un-directed graph like FaceBook (who-isFriendsWith-whom), with new nodes and edges added over time. The diameter $D(t)$ at month $t$ was 3, 5, 15, 9 ($t = 1, \ldots, 4$). According to the book and the lecture notes, what will the value of the diameter $D(5)$ be, on the next month ($t = 5$)?

- A. 15 *# it will go back to its highest value*
- B. between 9-15 *# it will almost go back, but a bit lower*
- C. 9 *# it will stay there*
- D. 3 *# it will keep on dropping with the same rate*
- E. between 5-9 *# it will keep dropping, approaching ≈6.*
- F. 2-3 *# it will densify - graph is clearly beyond 'gelling' point*
- G. something else - explain: .........

(b) **[3 points]** Consider a different graph, with $N$ isolated nodes (ie, $E$=0 edges). What is its diameter? (Reminder: the radius of a node is the distance to the most remote (but reachable) node; the diameter of a graph is the maximum radius over all nodes.)

(b) ____________

(c) **[3 points]** We have yet-another graph with $N'$ nodes. What is the smallest number of edges $E$ that will make the graph to have diameter $D$=1?

(c) ____________