Q12 — Mining Data Streams (12 points)

Setup

- Stream of ⟨user, topic⟩ tuples
- Set P of topics of interest
- Total platform users: 100,000
- **Algorithm:** (1) Bloom filter on topics → filter stream to P-related tuples, (2) Flajolet-Martin on users → count distinct users

Company policy (strict): OK to count some extra users (FP allowed), NOT OK to miss any user who posted about P (FN not allowed). This aligns perfectly with Bloom filters (no false negatives).

Part 1 (4 points) — Optimal k Range

k	FPR
3	0.95%
8	0.57%
12	0.32%
14	1.34%

The Bloom filter FPR curve $(1 - e^{-km/n})^k$ is convex: it decreases to a minimum at the optimal k, then increases.

Observations:

- FPR decreases monotonically from k=3 → 8 → 12 → optimal k ⩾ 12
- FPR increases from k=12 → 14 → optimal k < 14

$$\boxed{\text{Ranges (c) } 8 < k < 12 \text{ and (d) } 12 < k < 14}$$

Both ranges could contain the optimal k, since the minimum lies somewhere in [12, 14).

Part 2 (4 points) — Estimated Users Posting About P

Flajolet-Martin: max trailing zeros = 12 → raw estimate of distinct users passing Bloom filter = $2^{12} = 4096$.

This count includes **false positives** from the Bloom filter. Let x = true number of P-posting users.

The (100,000 − x) non-P users each pass the filter with probability 0.25% (FPR):

$$x + (100{,}000 - x) \times 0.0025 = 4096$$

$$x + 250 - 0.0025x = 4096$$

$$0.9975x = 3846$$

$$\boxed{x = \frac{4096 - 250}{1 - 0.0025} = \frac{3846}{0.9975} \approx 3856}$$

Part 3 (4 points) — Algorithm Under Relaxed Policy

New policy: Counting non-P users is NOT OK (no FP), but missing some P users is fine (FN allowed).

Answer: No, the algorithm no longer works.

The Bloom filter guarantees **no false negatives** but **allows false positives.** This was exactly what the old strict policy needed. Under the new policy, the requirements are reversed:

Property needed	Old (strict) policy	New (relaxed) policy
False negatives	Not OK ✓ (Bloom filter has none)	OK
False positives	OK ✓ (Bloom filter may have some)	Not OK ✗ (Bloom filter has some!)

The Bloom filter will let through users whose topics are NOT in P (false positives), directly violating the new policy. The algorithm's fundamental guarantee is mismatched with the new requirement.