CS246 Exam 2025 — Question 2: Locality Sensitive Hashing (14 points)

Part 1a (2 points) — Which curve corresponds to b = 2, r = 3?

The S-curve formula for the probability that two items with Jaccard similarity s become candidates is:

$$P(s) = 1 - (1 - s^r)^b$$

Since b × r = 6, there are four configurations. Each has a characteristic **threshold** ≈ (1/b)^(1/r) — the similarity where the curve transitions steeply:

Config	b	r	Threshold	P(0.5)	Curve
b=6, r=1	6	1	0.167	0.984	**(A)** — rises earliest
b=3, r=2	3	2	0.577	0.578	**(B)**
b=2, r=3	2	3	0.794	0.234	**(C)**
b=1, r=6	1	6	1.000	0.016	**(D)** — rises latest

Answer: (C) Option C

Intuition: Larger r (more rows per band) makes it harder for any band to match. Smaller b (fewer bands) gives fewer chances. Both effects push the threshold rightward.

Part 1b (2 points) — Fix r, increase b: does the false positive rate increase or decrease?

Answer: (A) Increase

From the formula $P(s) = 1 - (1 - s^r)^b$: as b grows, $(1 - s^r)^b$ shrinks toward 0, so P(s) increases for every s. The S-curve shifts left and upward, meaning more low-similarity pairs get flagged as candidates — these are false positives.

Each band is an independent "chance" for a pair to become a candidate. More bands = more chances = more items get through, including items that shouldn't.

Part 2a (1 point) — Jaccard similarity J(A, B)

A = {a, b, c, d, e, f, g, h}, B = {b, c, d, e, f, g, h, i, j}

- A ∩ B = {b, c, d, e, f, g, h} → 7 elements
- A ∪ B = {a, b, c, d, e, f, g, h, i, j} → 10 elements

$$\boxed{J(A, B) = 7/10 = 0.7}$$

Part 2b (1 point) — Jaccard similarity J(A, C)

A = {a, b, c, d, e, f, g, h}, C = {b, f, h, m, k}

- A ∩ C = {b, f, h} → 3 elements
- A ∪ C = {a, b, c, d, e, f, g, h, m, k} → 10 elements

$$\boxed{J(A, C) = 3/10 = 0.3}$$

Part 2c (2 points) — Optimal b and r for 6 MinHash functions

We need P(0.7) ⩾ 0.85 and P(0.3) ⩽ 0.25. Checking all configurations with b × r = 6:

Config	P(0.7)	P(0.3)	Both satisfied?
b=6, r=1	0.999 ✓	0.882 ✗	No — too many false positives
b=3, r=2	**0.867 ✓**	**0.246 ✓**	**Yes**
b=2, r=3	0.568 ✗	—	No — misses too many true positives
b=1, r=6	0.118 ✗	—	No

Verifying b=3, r=2:

- P(0.7) = 1 − (1 − 0.49)³ = 1 − (0.51)³ = 1 − 0.133 = **0.867** ⩾ 0.85 ✓
- P(0.3) = 1 − (1 − 0.09)³ = 1 − (0.91)³ = 1 − 0.754 = **0.246** ⩽ 0.25 ✓

$$\boxed{b = 3, \quad r = 2}$$

Part 3a (4 points) — Approximate errors for P_1 and P_2

Two constructions for similarity $x = 1 - \varepsilon$, where $\varepsilon \ll 1$:

- **P_1** (AND-then-OR): $P_1(x) = 1 - (1 - x^r)^b$
- **P_2** (OR-then-AND): $P_2(x) = (1 - (1 - x)^b)^r$

Using the approximation $(1 - x)^p \approx 1 - px$ for small x:

Error for P_1:

$$1 - P_1(x) = (1 - x^r)^b$$

Since $x = 1 - \varepsilon$: $x^r = (1-\varepsilon)^r \approx 1 - r\varepsilon$, so $1 - x^r \approx r\varepsilon$. Therefore:

$$\boxed{1 - P_1(x) \approx (r\varepsilon)^b = r^b \cdot \varepsilon^b}$$

Error for P_2:

$$1 - P_2(x) = 1 - (1 - (1 - x)^b)^r = 1 - (1 - \varepsilon^b)^r$$

Since ε^b is very small ($\varepsilon \ll 1$, $b \geqslant 2$), we apply the approximation: $(1 - \varepsilon^b)^r \approx 1 - r\varepsilon^b$. Therefore:

$$\boxed{1 - P_2(x) \approx r \cdot \varepsilon^b}$$

Part 3b (2 points) — Which construction to use?

We want the smallest error (i.e., highest $P(x)$):

Construction	Error
P_1 (AND-then-OR)	$\mathbf{r^b \cdot \varepsilon^b}$
P_2 (OR-then-AND)	$\mathbf{r \cdot \varepsilon^b}$

Since $r \geqslant 2$ and $b \geqslant 2$, we have $r^b > r$, so $1 - P_2(x) < 1 - P_1(x)$, meaning $P_2(x) > P_1(x)$.

$$\boxed{\textbf{(ii)}\ P_2 \text{ — the OR-then-AND construction}}$$

Intuition: For near-identical documents ($x \approx 1$), the OR step fires almost for free — nearly every hash function agrees, so at least one in each group will match. Applying OR first wastes nothing, then AND confirms consistency across groups. In P_1, the AND step runs first and unnecessarily reduces the already-high probability before OR can help.