L5: Locality Sensitive Hashing

Jeff M. Phillips

January 22, 2020

Family hash functions $\mathcal{H}$

$$\Pr_{h \in \mathcal{H}}[h(p) = h(q)] \approx sim(p, q)$$

=

Jaccard

Triangle

≈

Euclidean

(dot product)

1. 1 hash function

$$\hat{JS}(p,q) = \begin{cases} 1 & h(p) = h(q) \\ 0 & o.w. \end{cases}$$

hash table

2. k hash functions

$$\hat{JS}_k(p,q) = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}(h_j(p) = h_j(q))$$

Algo?

Indicator Function

$$\mathbb{1}(b) = \begin{cases} 1 & \text{if } b = \text{True} \\ 0 & \text{if } b = \text{False} \end{cases}$$

Large Number of objects X

$$X = \{x_1, x_2, \ldots x_n\}$$

(documents (k-grams), IP addresses, customers)

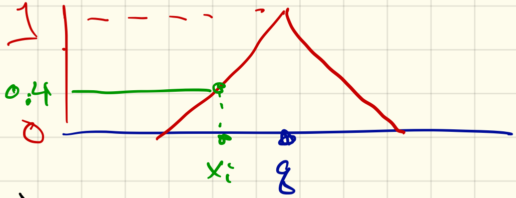
Q1 : Which pairs are similar?

n^2 time

Q2 : Given query q, which $x_i \in X$ are similar to q?

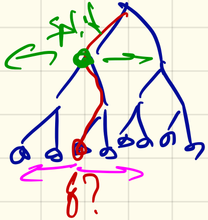
n time

$x_1, x_2, \dots x_n \in \mathbb{R}$



similarity $S_\Delta(q, x_i) = \max\{0, 1-|q-x_i|\}$

1. Sort $x_1, x_2 \dots$
2. Build binary tree T
3. Find q in T



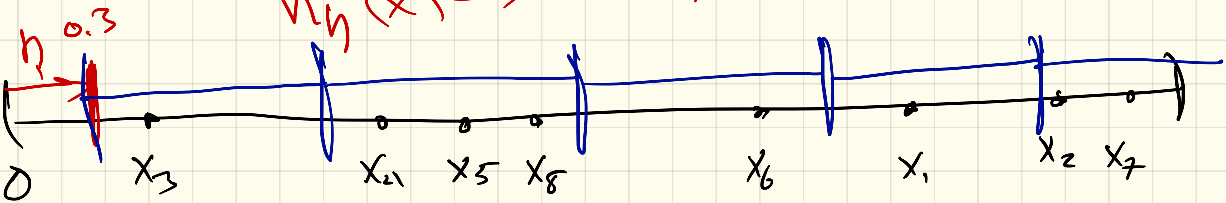
$\Pr[h(x)=h(x')] = S_\Delta(x, x')$

$\log n + k$

\# similar items

$\eta \in \mathrm{unif}(0,1)$

$h_\eta(x) \to$ a bin



Banding : How to combine hash functions

$H = \{h_1, h_2, \dots h_t\} \in \mathcal{H}$

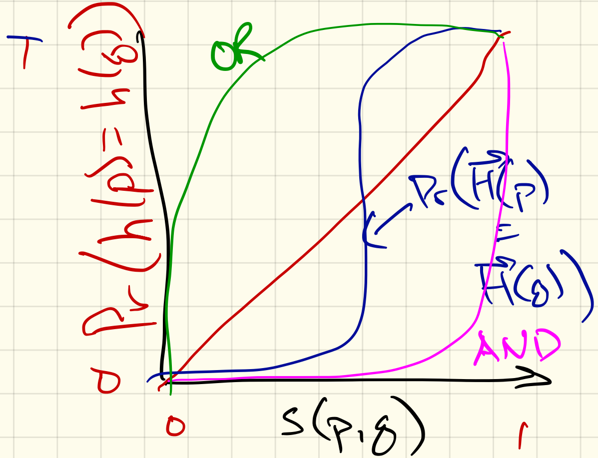
$\vec{H} \leftarrow$ single super hash function

band $b = 2$

	h_1	h_2	h_3	h_4	h_5	h_6
p_1	3	5	0	2	4	3
p_2	3	4	1	0	2	2
p_3	0	2	4	3	5	1
p_4						
p_5						
p_6						
q	3	3	1	2	3	2

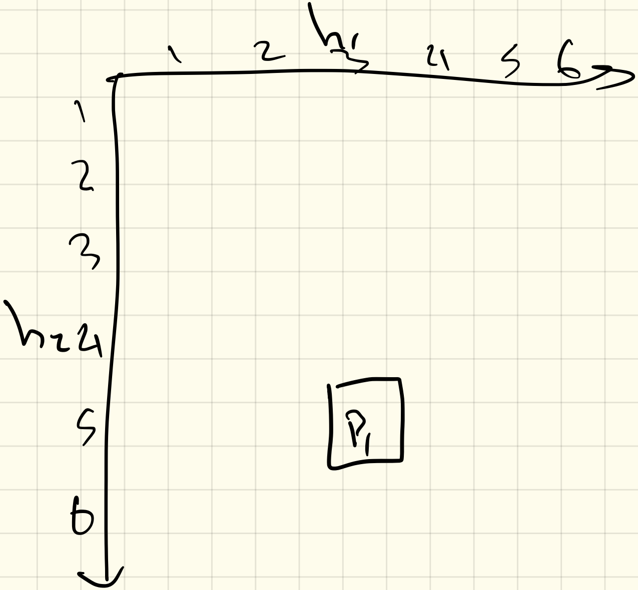
H_1 H_2 H_3

#bands $= r$



$$\vec{H}(p,q) = OR(H_1, H_2, H_3)$$

$P_1 \to [3, 5]$



$$Pr[h_1 h_2(p) = h_1 h_2(q)]$$

$$= s^2$$

Much more selective

r bands, each with b hash functions

$t = \#$ hash functions $\quad t \geq r \cdot b$

$s(p, q) = s$

$s^b = \Pr$ p, q collide in one band.

$(1 - s^b) = \Pr$ p, q don't collide

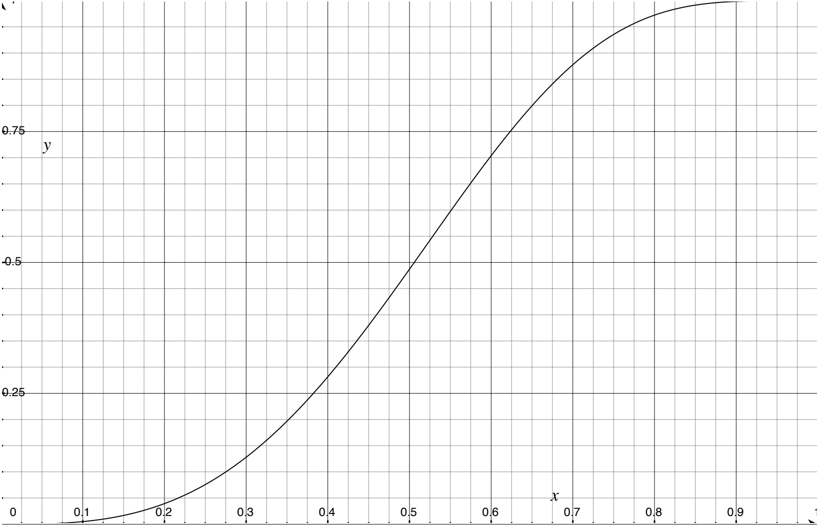
$(1 - s^b)^r = \Pr$ p, q don't collide in r bands

$f(s) = 1 - (1 - s^b)^r = \Pr$ p, q collide in at least one band.

LSH $b = 3$ and $r = 5$

$t = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

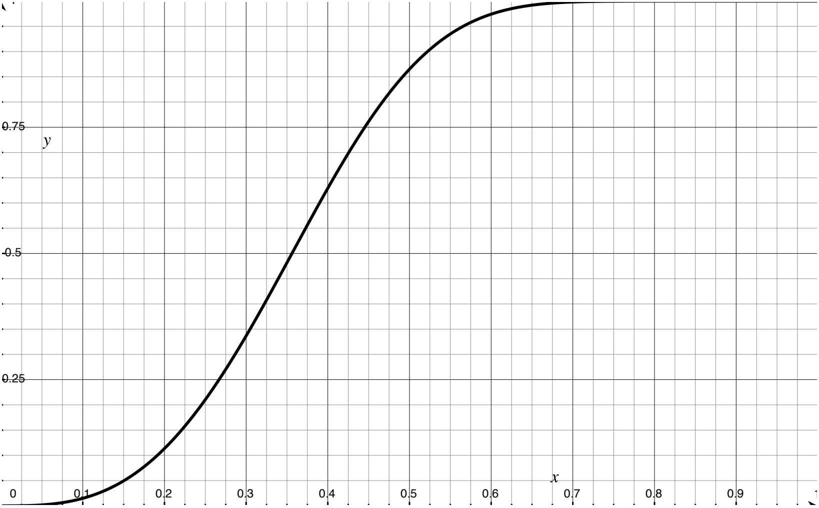


LSH $b = 3$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

LSH $b = 3$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$



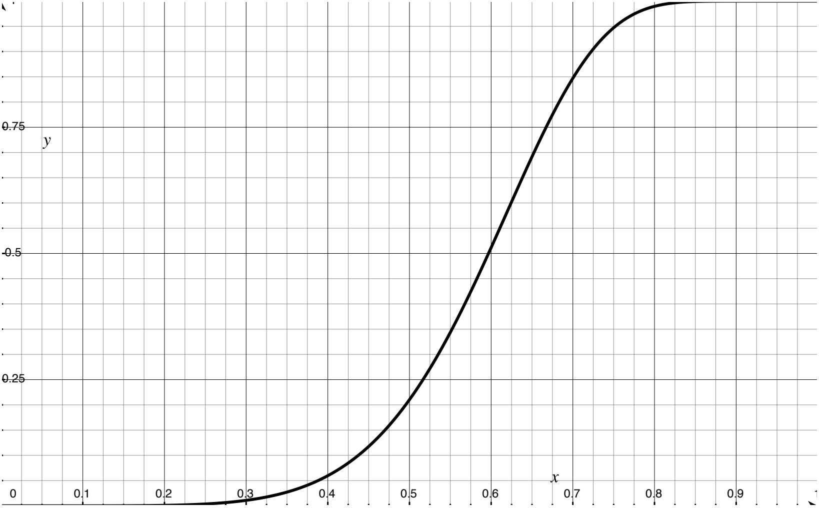
LSH $b = 6$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

LSH $b = 6$ and $r = 15$

$t = 90$

Probability of found collision $= 1 - (1 - s^b)^r$

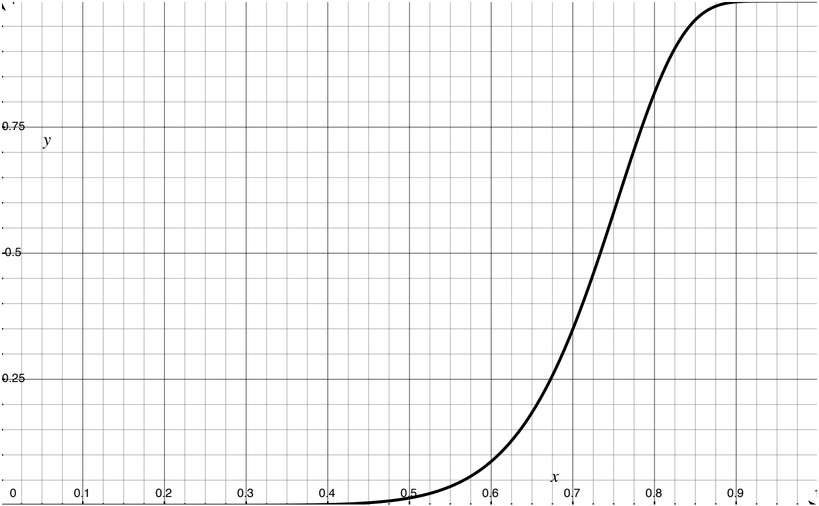


LSH $b = 10$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

LSH $b = 10$ and $r = 15$

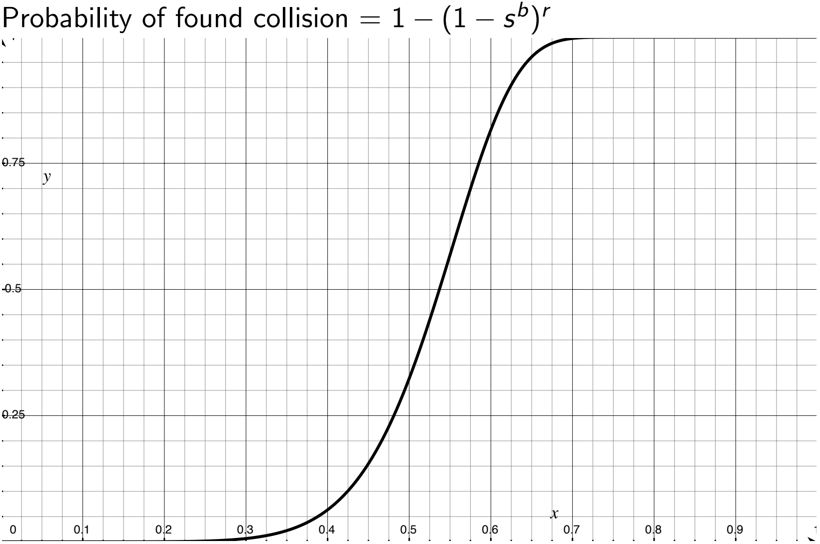
Probability of found collision $= 1 - (1 - s^b)^r$



LSH $b = 8$ and $r = 100$

Probability of found collision $= 1 - (1 - s^b)^r$

LSH $b = 8$ and $r = 100$

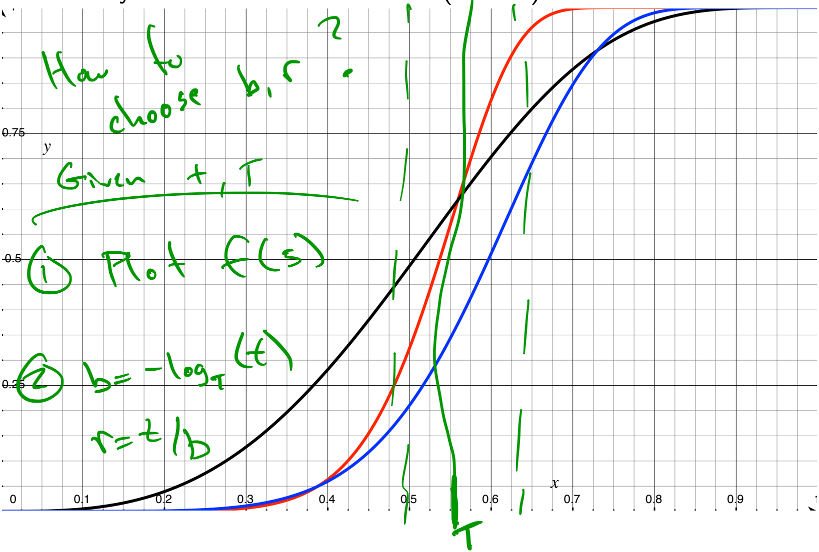


LSH $(b = 3, r = 5)$ & $(b = 6, r = 15)$ & $(b = 8, r = 100)$

Probability of found collision $= 1 - (1 - s^b)^r$

LSH $(b = 3, r = 5)$ & $(b = 6, r = 15)$ & $(b = 8, r = 100)$

Probability of found collision $= 1 - (1 - s^b)^r$



LSH for Euclidean Dist.

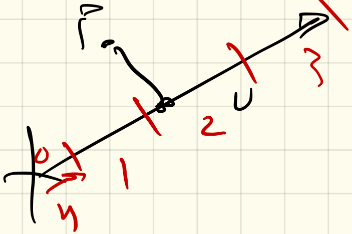
$$d_E(p,q) \Longleftrightarrow s_E(p,q)$$

$$\langle p, q \rangle = \sum_{j=1}^{d} p_i \cdot q_i$$

$$h_{\eta,u} : \mathbb{R}^d \to [m]$$

$$\eta \in \text{unif}(0,T)$$

$$u \in \mathbb{R}^d, \ \|u\| = 1$$



$$h_{\eta,u}(p) = \left(\underbrace{\lfloor \langle p, u \rangle - \eta \rfloor}_{\text{floor}} \bmod m \right)$$