Locality-Sensitive Hashing (LSH)

Additional Materials

Mining Massive Datasets

Materials provided by Prof. Carlos Castillo — https://chato.cl/teach

Instructor: Dr. Teodora Sandra Buda — https://tbuda.github.io/

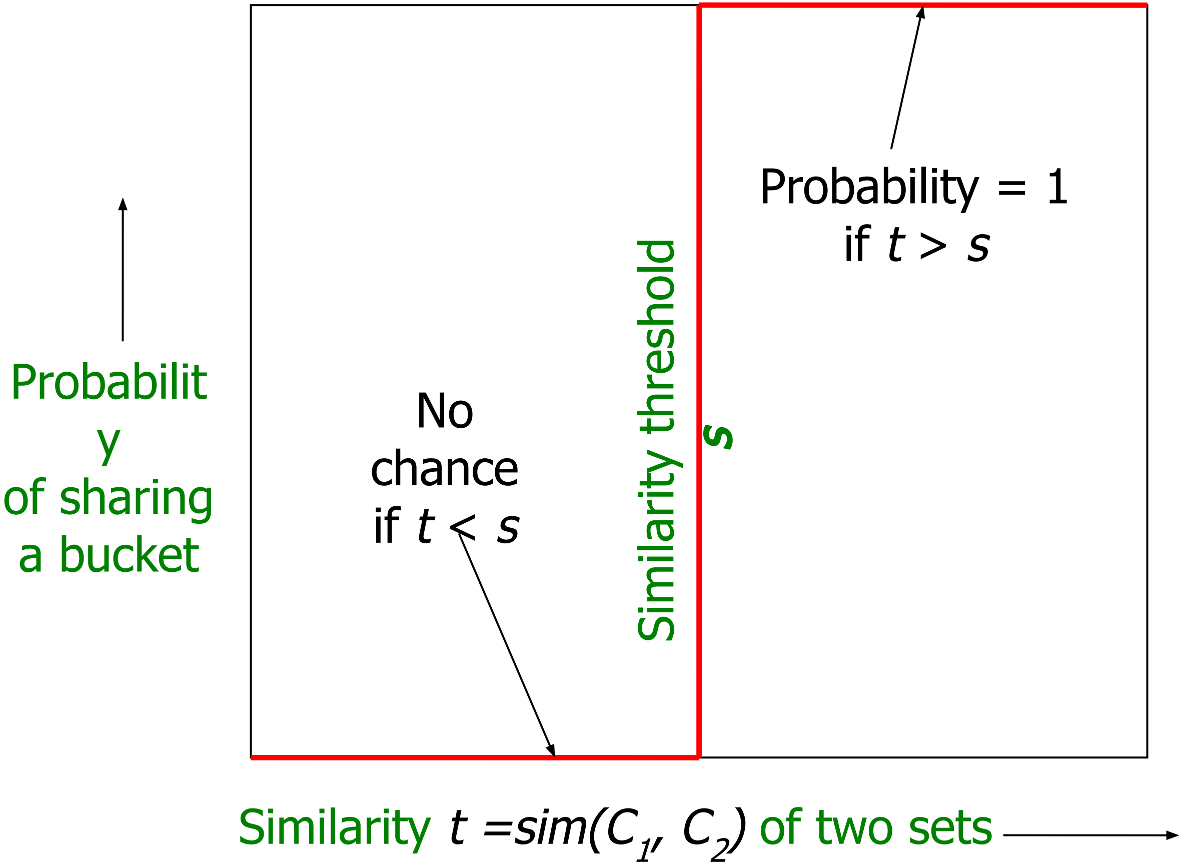
Source for this deck

- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. (Chapter 3) [slides ch3]

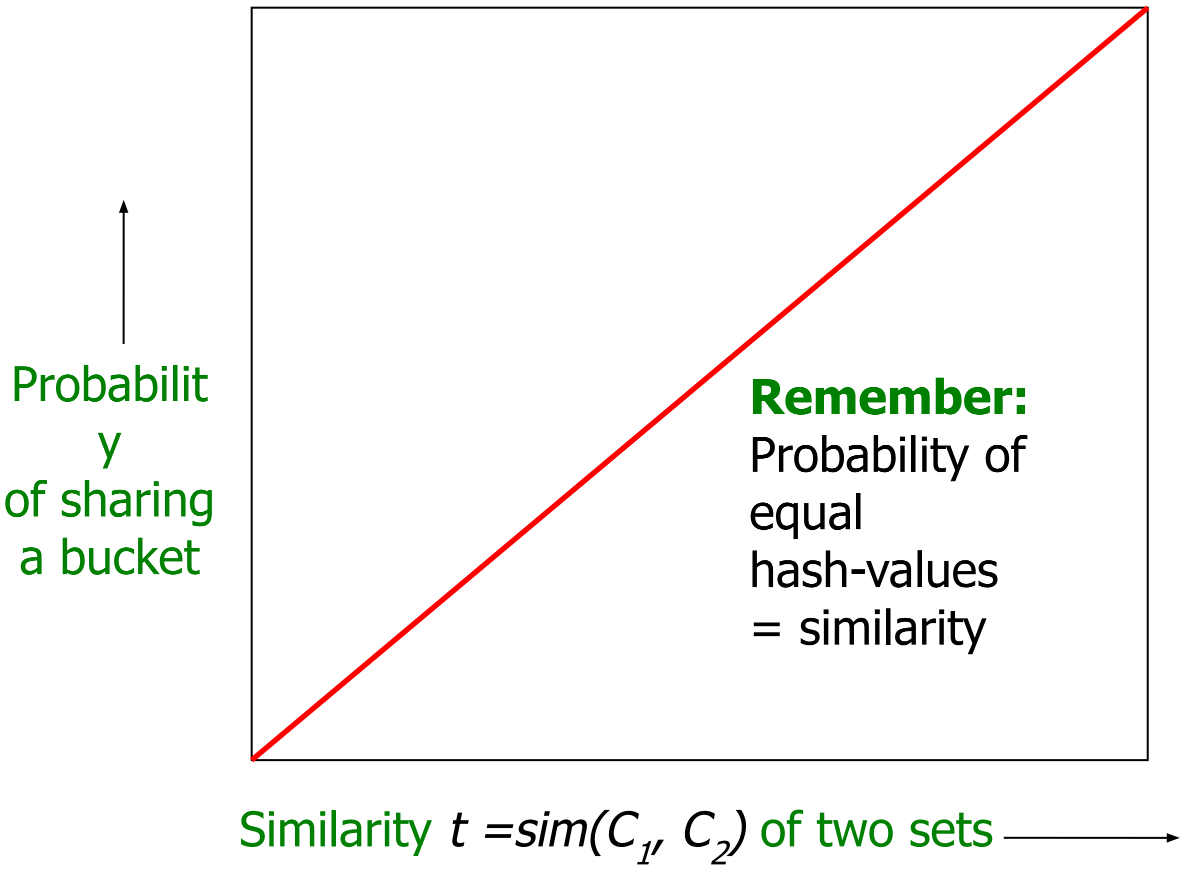
LSH involves a trade-off

- Pick:
  - The number of Min-Hashes (rows of M = K)
  - The number of bands b, and
  - The number of rows r per band to balance false positives/negatives
- Example: If we had only 15 bands of 5 rows, the number of false positives would go down, but the number of false negatives would go up

LSH: what we want



What 1 band of 1 row gives you

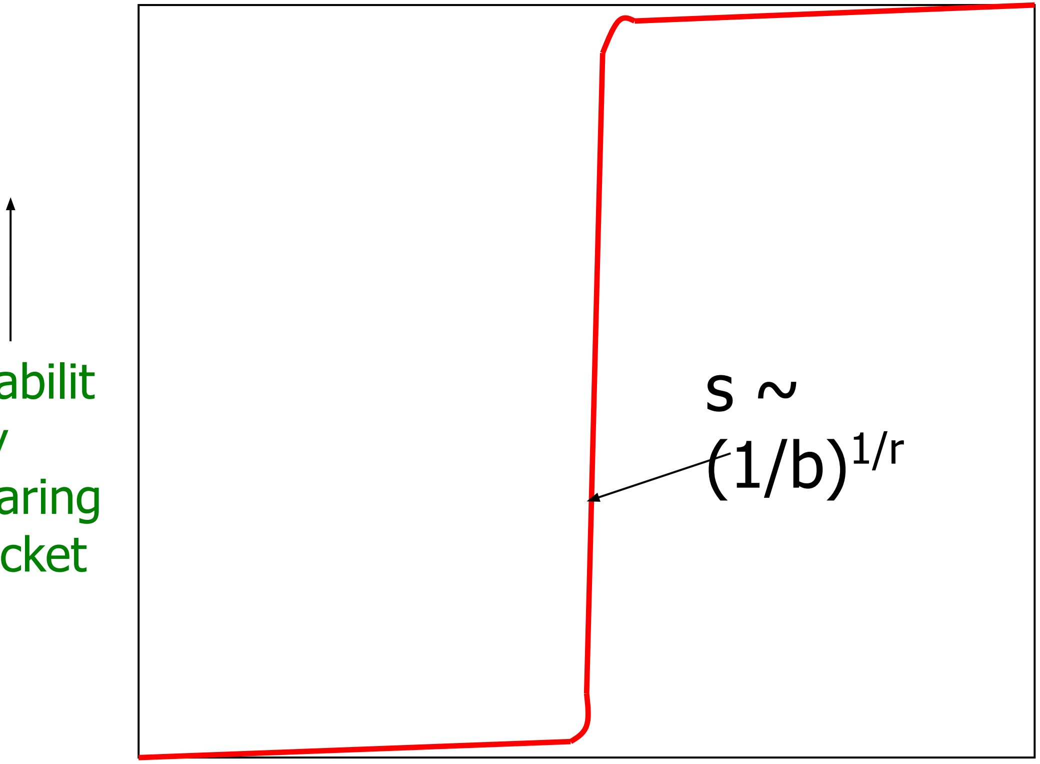


Remember:
Probability of equal hash-values = similarity

b bands, r rows/band

- Columns C_1 and C_2 have similarity ***t***
- Pick any band (***r*** rows)
  - Prob. that all rows in band equal = t^r
  - Prob. that some row in band unequal = $1 - t^r$
- Prob. that no band identical = $(1 - t^r)^b$
- Prob. that at least 1 band identical = $1 - (1 - t^r)^b$

What b bands of r rows give you



Probability of sharing a bucket

s ~ $(1/b)^{1/r}$

Similarity $t=sim(C_1, C_2)$ of two sets

At least one band identical

No bands identical

$$1 - (1 - t^r)^b$$

Some row of a band unequal

All rows of a band are equal

Example: b=20, r=5

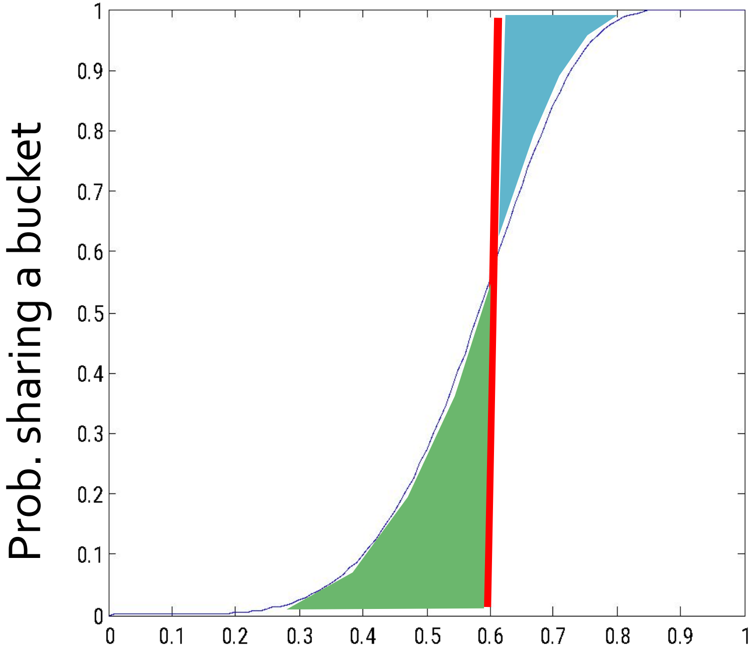
- Similarity threshold s
- Prob. that at least 1 band is identical:

s	$1-(1-s^r)^b$
.2	.006
.3	.047
.4	.186
.5	.470
.6	.802
.7	.975
.8	.9996

Picking r and b: the S curve

Picking *r* and *b* to get the best S-curve

50 hash-functions (r=5, b=10)



Blue area: False Negative rate

Green area: False Positive rate

Summary

Things to remember

- **Locality-Sensitive Hashing:** Focus on pairs of signatures likely to be from similar documents
  - We used hashing to find **candidate pairs** of similarity ≥ **s**

Exercises for TT08-TT09

- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al.
  - Exercises 3.1.4 (Jaccard similarity)
  - Exercises 3.2.5 (Shingling)
  - Exercises 3.3.6 (Min hashing)
  - Exercises 3.4.4 (Locality-sensitive hashing)