Locality-Sensitive Hashing

Focusing on Similar Minhash Signatures
Other Applications Will Follow

Locality-Sensitive Hashing

- General idea: Generate from the collection of all elements (signatures in our example) a small list of *candidate pairs*: pairs of elements whose similarity must be evaluated.
- For signature matrices: Hash columns to many buckets, and make elements of the same bucket candidate pairs.

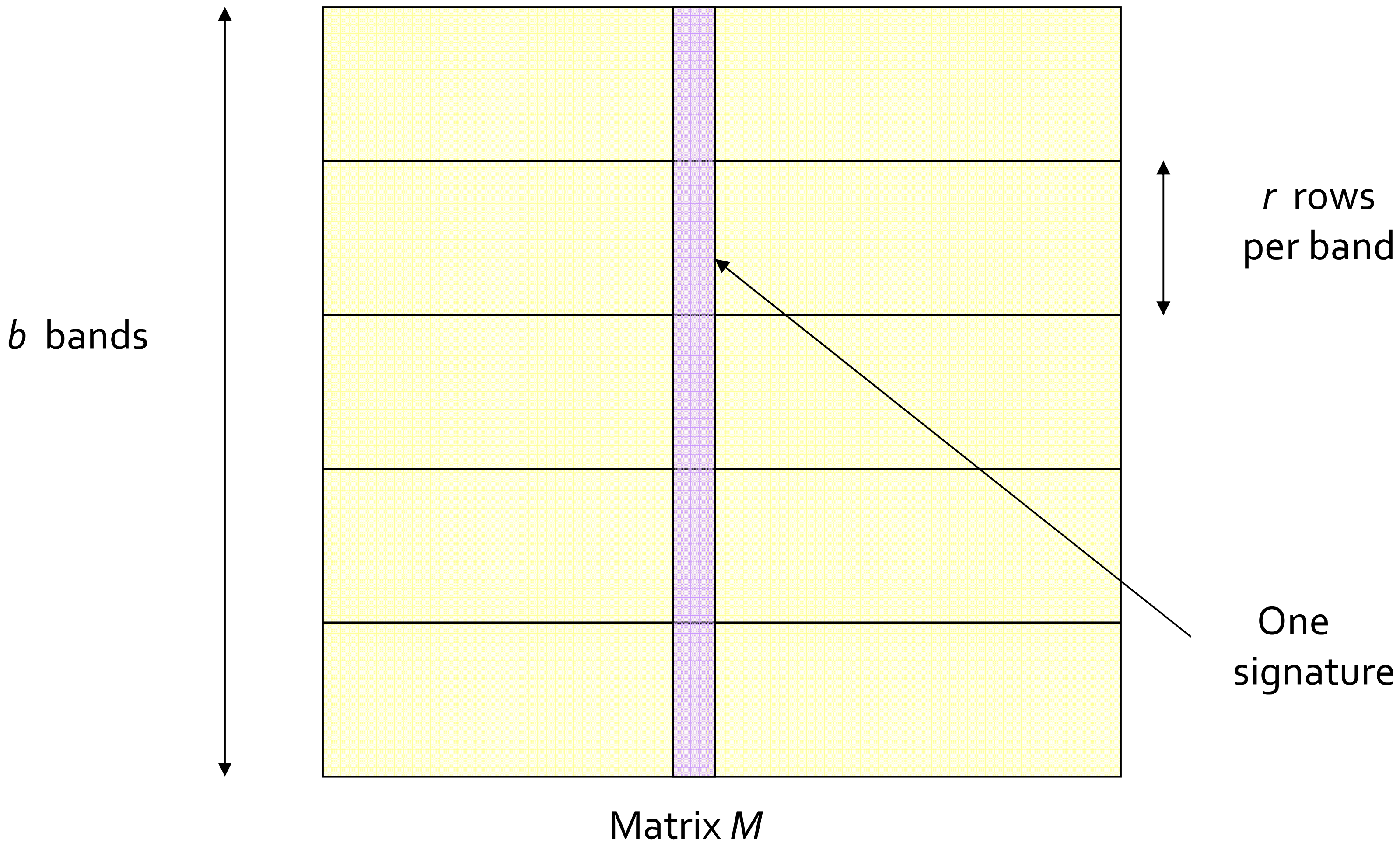
Candidate Generation From Minhash Signatures

- Pick a similarity threshold *t*, a fraction < 1.
- We want a pair of columns *c* and *d* of the signature matrix M to be a *candidate pair* if and only if their signatures agree in at least fraction *t* of the rows.
  - I.e., $M(i, c) = M(i, d)$ for at least fraction *t* values of *i*.

LSH for Minhash Signatures

- Big idea: hash columns of signature matrix *M* several times.
- Arrange that (only) similar columns are likely to hash to the same bucket.
- Candidate pairs are those that hash at least once to the same bucket.

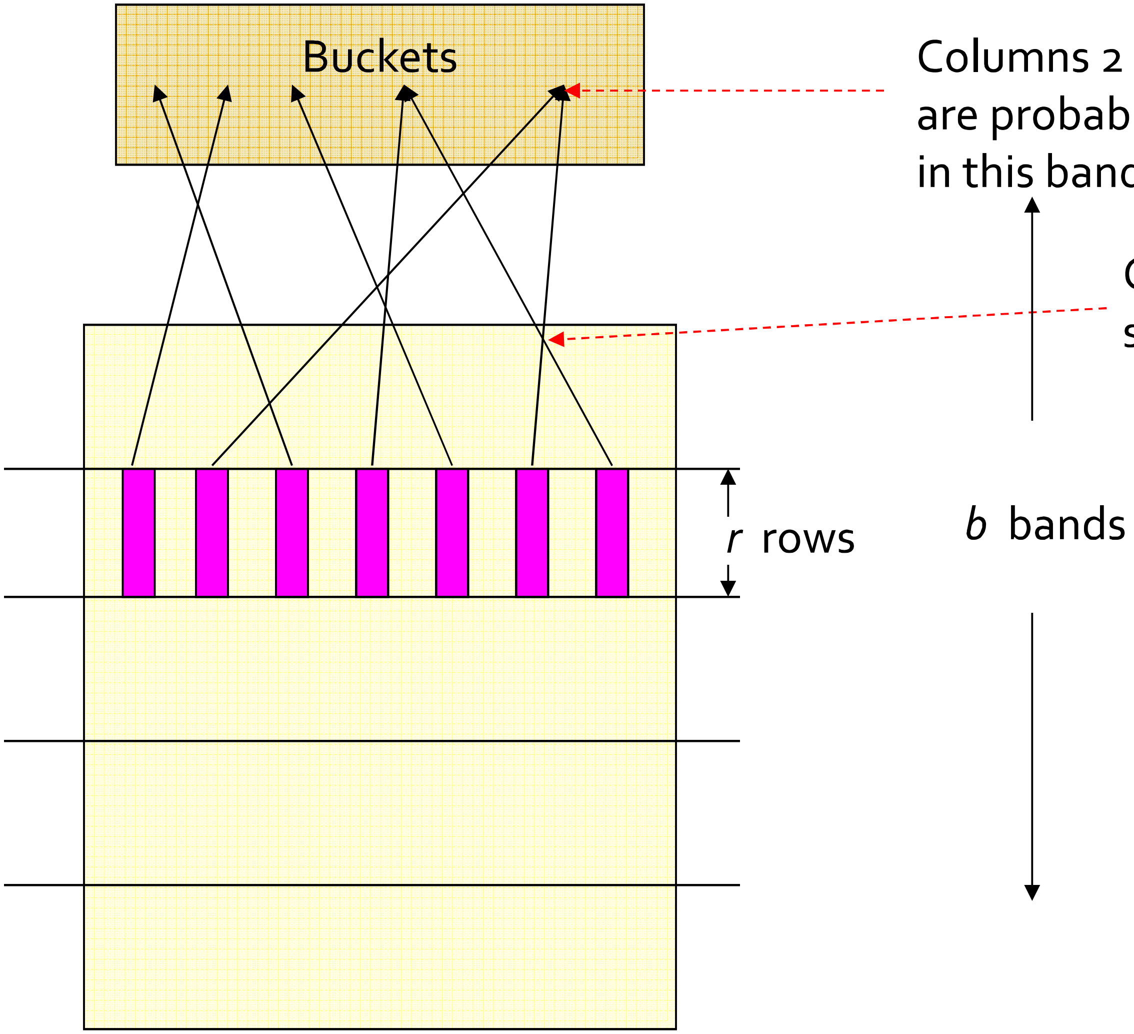
Partition Into Bands



Partition into Bands – (2)

- Divide matrix *M* into *b* bands of *r* rows.
- For each band, hash its portion of each column to a hash table with *k* buckets.
  - Make *k* as large as possible.
- *Candidate* column pairs are those that hash to the same bucket for ≥ 1 band.
- Tune *b* and *r* to catch most similar pairs, but few nonsimilar pairs.

Hash Function for One Bucket



Columns 2 and 6 are probably identical in this band.

Columns 6 and 7 are surely different.

Matrix M

Example – Bands

- Suppose 100,000 columns.
- Signatures of 100 integers.
- Therefore, signatures take 40Mb.
- Want all 80%-similar pairs of documents.
- 5,000,000,000 pairs of signatures can take a while to compare.
- Choose 20 bands of 5 integers/band.

Suppose C_1, C_2 are 80% Similar

- Probability C_1, C_2 identical in one particular band: $(0.8)^5 = 0.328$.
- Probability C_1, C_2 are *not* similar in any of the 20 bands: $(1-0.328)^{20} = .00035$.
  - i.e., about 1/3000th of the 80%-similar underlying sets are false negatives.