Review: Sketches

How do we save a small number of bits about things so we can store them?

Sketch - a few hundred bits that allow us to answer questions about full data.

Example: Each student selects 1000 integers in range 1 to 10⁶. Create sketch consisting of 10 integers.

Define: resemblance:

$$r(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

1. Select 10 elements at random from each set. Won’t work - need \(\sqrt{n}\) to get any overlap

\(\Rightarrow\) Birthday paradox.

2. Sort sets and select 10 minimum integers.

$$m(A) = 10$$ minimum of \(A\)

$$r(A, B) = \frac{m(A) \cap m(B)}{m(A) \cup m(B)} = \frac{5}{15} = 0.3$$

*Note: we pick a random sort order.

Could have done: Select elements equal to \(0 \mod m\)

Could ask: is \(A \subseteq B\)

Want to extend to sequences

Look at k-shingles: a consecutive sequence of k symbols.

If k is large enough, from set of k-shingles, I can reconstruct the sequence.

Consider sequence of length 10,000 over alphabet of size 100.
Sequence has $10^4$ shingles out of universe of size $10^k$.

select $k=3$ → now universe is of size $10^6$ 100 possible successor shingles

What is the probability that a given shingle has two or more successor shingles in the set of shingles of our sequence

$$P(\text{not failure}) \left(1 - \frac{1}{100}\right)^{100} \approx \frac{1}{e} = 0.6$$

select $k=5$ → universe is of size $10^6$

$$\frac{10^4}{10^{10}} = \frac{1}{10^6} \quad P(\text{not failure}) \left(1 - \frac{10^4}{10^{10}}\right)^{100} = \left(1 - \frac{1}{10^6}\right)^{100} = 0.9999$$

(for entire sequence) $\left(\left(1 - \frac{1}{10^6}\right)^{100}\right)^{10000} = \frac{1}{e}$