

# COMPSCI 514: Algorithms for Data Science

---

Cameron Musco

University of Massachusetts Amherst. Fall 2024.

Lecture 6

- Problem Set 1 is due tomorrow at 11:59pm in Gradescope.
- Remember to make one submission per group, but all group members must be enrolled in Gradescope and have their names added to the submission.
- Submit 4 of the 5 questions and just submit nothing for the challenge problem that you didn't complete.
- Quiz 3 is due Monday at 8pm.

# Last Time

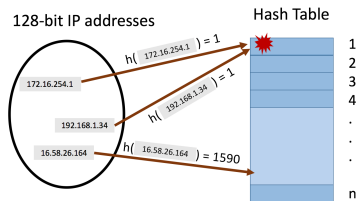
## Last Class:

- Higher moment bounds and exponential concentration bounds
- Bernstein inequality and the Chernoff bound
- Connection to the central limit theorem.

## This Class:

- Finish example application of exponential concentration bounds.
- Bloom filters: random hashing to maintain a large set in small space.

# Application to Random Hashing



We hash  $m$  values  $x_1, \dots, x_m$  using a random hash function into a table with  $n = m$  entries.

- I.e., for all  $j \in [m]$  and  $i \in [m]$ ,  $\Pr(\mathbf{h}(x_j) = i) = \frac{1}{m}$  and hash values are chosen independently.

What will be the maximum number of items hashed into the same location?

# Maximum Load in Randomized Hashing

Let  $S_i$  be the number of items hashed into position  $i$  and  $S_{i,j}$  be 1 if  $x_j$  is hashed into bucket  $i$  ( $h(x_j) = i$ ) and 0 otherwise.

$$\mathbb{E}[S_i] = \sum_{j=1}^m \mathbb{E}[S_{i,j}] = m \cdot \frac{1}{m} = 1$$

**By the Chernoff Bound:** for any  $\delta \geq 0$ ,

$$\Pr(S_i \geq 1 + \delta) \leq \Pr\left(\left|\sum_{i=1}^n S_{i,j} - 1\right| \geq \delta \cdot \mu\right) \leq 2 \exp\left(-\frac{\delta^2}{2 + \delta}\right)$$

$m$ : total number of items hashed and size of hash table.  $x_1, \dots, x_m$ : the items.  
 $h$ : random hash function mapping  $x_1, \dots, x_m \rightarrow [m]$ .

## Maximum Load in Randomized Hashing

$$\Pr(S_i \geq 1 + \delta) \leq \Pr\left(\left|\sum_{j=1}^n S_{i,j} - 1\right| \geq \delta\right) \leq 2 \exp\left(-\frac{\delta^2}{2 + \delta}\right).$$

Set  $\delta = 5 \log m$ . Gives:

$$\Pr(S_i \geq 5 \log m + 1) \leq 2 \exp\left(-\frac{(5 \log m)^2}{2 + 5 \log m}\right) \leq 2 \exp(-3 \log m) \leq \frac{2}{m^3}.$$

Apply Union Bound:

$$\begin{aligned}\Pr(\max_{i \in [m]} S_i \geq 5 \log m + 1) &= \Pr\left(\bigcup_{i=1}^m (S_i \geq 5 \log m + 1)\right) \\ &\leq \sum_{i=1}^m \Pr(S_i \geq 5 \log m + 1) \leq m \cdot \frac{2}{m^3} = \frac{2}{m^2}.\end{aligned}$$

$m$ : total number of items hashed and size of hash table.  $S_i$ : number of items hashed to bucket  $i$ .  $S_{i,j}$ : indicator if  $x_j$  is hashed to bucket  $i$ .  $\delta$ : any value  $\geq 0$ .

## Maximum Load in Randomized Hashing

**Upshot:** If we randomly hash  $m$  items into a hash table with  $m$  entries the maximum load per bucket is  $O(\log m)$  with very high probability.

- So, even with a simple linked list to store the items in each bucket, worst case query time is  $O(\log m)$ .
- Using Chebyshev's inequality could only show the maximum load is bounded by  $O(\sqrt{m})$  with good probability (good exercise).
- The Chebyshev bound holds even with a pairwise independent hash function. The stronger Chernoff-based bound can be shown to hold with a *k-wise independent hash function* for  $k = O(\log m)$ .

## Approximately Maintaining a Set

Want to store a set  $S$  of items from a massive universe of possible items (e.g., images, text documents, IP addresses).

**Goal:** support  $insert(x)$  to add  $x$  to the set and  $query(x)$  to check if  $x$  is in the set. Both in  $O(1)$  time. **What data structure solves this problem?**

- Allow small probability  $\delta > 0$  of false positives. I.e., for any  $x$ ,

$$\Pr(query(x) = 1 \text{ and } x \notin S) \leq \delta.$$

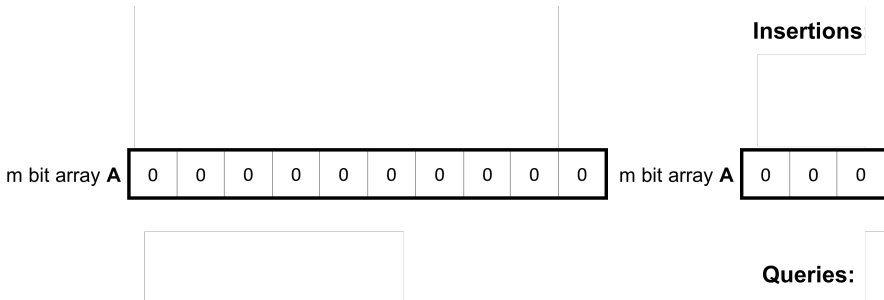
**Solution:** Bloom filters (repeated random hashing). Will use much less space than a hash table.



# Bloom Filters

Chose  $k$  independent random hash functions  $h_1, \dots, h_k$  mapping the universe of elements  $U \rightarrow [m]$ .

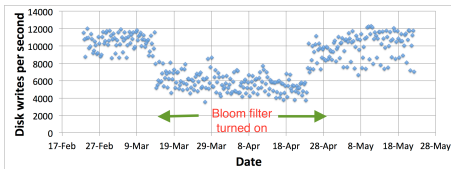
- Maintain an array  $A$  containing  $m$  bits, all initially 0.
- *insert*( $x$ ): set all bits  $A[h_1(x)] = \dots = A[h_k(x)] := 1$ .
- *query*( $x$ ): return 1 only if  $A[h_1(x)] = \dots = A[h_k(x)] = 1$ .



No false negatives. False positives more likely with more insertions.

# Applications: Caching

Akamai (Boston-based company serving 15 – 30% of all web traffic) applies bloom filters to prevent caching of ‘one-hit-wonders’ – pages only visited once fill over 75% of cache.



- When url  $x$  comes in, if  $query(x) = 1$ , cache the page at  $x$ . If not, run  $insert(x)$  so that if it comes in again, it will be cached.
- **False positive:** A new url (possible one-hit-wonder) is cached. If the bloom filter has a false positive rate of  $\delta = .05$ , the number of cached one-hit-wonders will be reduced by at least 95%.

# Applications: Databases

Distributed database systems, including Google Bigtable, Apache HBase, Apache Cassandra, and PostgreSQL use bloom filters to prevent expensive lookups of non-existent data.

Movies

	5			1	4				
		3						5	
Users					4				
		5							5
	1			2					

- When a new rating is inserted for  $(user_x, movie_y)$ , add  $(user_x, movie_y)$  to a bloom filter.
- Before reading  $(user_x, movie_y)$  (possibly via an out of memory access), check the bloom filter, which is stored in memory.
- **False positive:** A read is made to a possibly empty cell. A  $\delta = .05$  false positive rate gives a 95% reduction in these empty reads.

## More Applications

- **Database Joins:** Quickly eliminate most keys in one column that don't correspond to keys in another.
- **Recommendation systems:** Bloom filters are used to prevent showing users the same recommendations twice.
- **Spam/Fraud Detection:**
  - Bit.ly and Google Chrome use bloom filters to quickly check if a url maps to a flagged site and prevent a user from following it.
  - Can be used to detect repeat clicks on the same ad from a single IP-address, which may be the result of fraud.
- **Digital Currency:** Some Bitcoin clients use bloom filters to quickly pare down the full transaction log to transactions involving bitcoin addresses that are relevant to them (SPV: simplified payment verification).

# Analysis

For a bloom filter with  $m$  bits and  $k$  hash functions, the insertion and query time is  $O(k)$ . How does the false positive rate  $\delta$  depend on  $m$ ,  $k$ , and the number of items inserted?

**Step 1:** What is the probability that after inserting  $n$  elements, the  $i^{\text{th}}$  bit of the array  $A$  is still 0?  $n \times k$  total hashes must not hit bit  $i$ .

$$\begin{aligned}\Pr(A[i] = 0) &= \Pr(\mathbf{h}_1(x_1) \neq i \cap \dots \cap \mathbf{h}_k(x_k) \neq i \\ &\quad \cap \mathbf{h}_1(x_2) \neq i \dots \cap \mathbf{h}_k(x_2) \neq i \cap \dots) \\ &= \underbrace{\Pr(\mathbf{h}_1(x_1) \neq i) \times \dots \times \Pr(\mathbf{h}_k(x_1) \neq i) \times \Pr(\mathbf{h}_1(x_2) \neq i) \dots}_{k \cdot n \text{ events each occurring with probability } 1-1/m} \\ &= \left(1 - \frac{1}{m}\right)^{kn}\end{aligned}$$

# Analysis

How does the false positive rate  $\delta$  depend on  $m$ ,  $k$ , and the number of items inserted?

**Step 1:** What is the probability that after inserting  $n$  elements, the  $i^{\text{th}}$  bit of the array  $A$  is still 0?

$$\Pr(A[i] = 0) = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-\frac{kn}{m}}$$

**Step 2:** What is the probability that querying a new item  $w$  gives a false positive?

$$\begin{aligned}\Pr(A[\mathbf{h}_1(w)] = \dots = A[\mathbf{h}_k(w)] = 1) \\ &= \Pr(A[\mathbf{h}_1(w)] = 1) \times \dots \times \Pr(A[\mathbf{h}_k(w)] = 1) \\ &= \left(1 - e^{-\frac{kn}{m}}\right)^k \quad \text{Actually Incorrect! Dependent events.}\end{aligned}$$

$n$ : total number items in filter,  $m$ : number of bits in filter,  $k$ : number of random hash functions,  $\mathbf{h}_1, \dots, \mathbf{h}_k$ : hash functions,  $A$ : bit array,  $\delta$ : false positive rate.

## Correct Analysis Sketch

**Step 1:** To avoid dependence issues, condition on the event that the  $A$  has  $t$  zeros in it after  $n$  insertions, for some  $t \leq m$ . For a non-inserted element  $w$ , after conditioning on this event we correctly have:

$$\begin{aligned}\Pr(A[\mathbf{h}_1(w)] = 1) &= \dots = \Pr(A[\mathbf{h}_k(w)] = 1) \\ &= \Pr(A[\mathbf{h}_1(w)] = 1) \times \dots \times \Pr(A[\mathbf{h}_k(w)] = 1).\end{aligned}$$

I.e., the events  $A[\mathbf{h}_1(w)] = 1, \dots, A[\mathbf{h}_k(w)] = 1$  are independent conditioned on the number of bits set in  $A$ . **Why?**

- Conditioned on this event, for any  $j$ , since  $\mathbf{h}_j$  is a fully random hash function,  $\Pr(A[\mathbf{h}_j(w)] = 1) = 1 - \frac{t}{m}$ .
- Thus conditioned on this event, the false positive rate is  $(1 - \frac{t}{m})^k$ .
- It remains to show that  $\frac{t}{m} \approx e^{-\frac{kn}{m}}$  with high probability. We already have that  $\mathbb{E}[\frac{t}{m}] = \frac{1}{m} \sum_{i=1}^m \Pr(A[i] = 0) \approx e^{-\frac{kn}{m}}$ .

## Correct Analysis Sketch

Need to show that the number of zeros  $t$  in  $A$  after  $n$  insertions is bounded by  $O\left(e^{-\frac{kn}{m}}\right)$  with high probability.

Can apply Theorem 2 of:

<http://cglab.ca/~morin/publications/ds/bloom-submitted.pdf>