COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Spring 2020.

Lecture 4

Last Class:

Application to bounding the maximum server load when using randomized routing.

Last Class:

- **Union bound:** $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ for *any* events A, B. Application to bounding the maximum server load when using randomized routing.

Last Class:

- **Union bound:** $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ for *any* events A, B. Application to bounding the maximum server load when using randomized routing.
- **From Markov's inequality to exponential tail bounds.**
- Bernstein, Chernoff bounds and the Central Limit Theorem.

Last Class:

- **Union bound:** $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ for *any* events A, B. Application to bounding the maximum server load when using randomized routing.
- **From Markov's inequality to exponential tail bounds.**
- Bernstein, Chernoff bounds and the Central Limit Theorem.

This Time:

Last Class:

- **Union bound:** $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ for *any* events A, B. Application to bounding the maximum server load when using randomized routing.
- **From Markov's inequality to exponential tail bounds.**
- Bernstein, Chernoff bounds and the Central Limit Theorem.

This Time:

- **Bloom Filters:** Random hashing to maintain a large sets in very small space.

Last Class:

- **Union bound:** $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ for *any* events A, B. Application to bounding the maximum server load when using randomized routing.
- **From Markov's inequality to exponential tail bounds.**
- Bernstein, Chernoff bounds and the Central Limit Theorem.

This Time:

- **Bloom Filters:** Random hashing to maintain a large sets in very small space.
- **Distinct Elements:** Estimating the number of unique items in a data stream via hashing. Prelude to audio fingerprinting, document search, etc.

A GENERAL COMMENT

We have covered a lot in the first three classes.

A GENERAL COMMENT

We have covered a lot in the first three classes.

The proofs in class are meant to illustrate techniques that can be used to tackle many algorithmic and data related problems. You do not need to have these proofs or their conclusions memorized.

A GENERAL COMMENT

We have covered a lot in the first three classes.

The proofs in class are meant to illustrate techniques that can be used to tackle many algorithmic and data related problems. You do not need to have these proofs or their conclusions memorized.

- Know and be comfortable applying: independence, linearity of expectation, linearity of variance, union bound, Markov's inequality, Chebyshev's inequality, basic probability calculations.

A GENERAL COMMENT

We have covered a lot in the first three classes.

The proofs in class are meant to illustrate techniques that can be used to tackle many algorithmic and data related problems. You do not need to have these proofs or their conclusions memorized.

- Know and be comfortable applying: independence, linearity of expectation, linearity of variance, union bound, Markov's inequality, Chebyshev's inequality, basic probability calculations.
- Able to use techniques like: breaking random variables into sums of indicator variables and analyzing expected collisions (used in CATCHA analysis, two-level hashing analysis, *and* variance calculations for randomized load balancing.)

A GENERAL COMMENT

We have covered a lot in the first three classes.

The proofs in class are meant to illustrate techniques that can be used to tackle many algorithmic and data related problems. You do not need to have these proofs or their conclusions memorized.

- Know and be comfortable applying: independence, linearity of expectation, linearity of variance, union bound, Markov's inequality, Chebyshev's inequality, basic probability calculations.
- Able to use techniques like: breaking random variables into sums of indicator variables and analyzing expected collisions (used in CATCHA analysis, two-level hashing analysis, *and* variance calculations for randomized load balancing.)
- Know definitions of 2-universal and pairwise independent hash functions and why they are useful.

A GENERAL COMMENT

We have covered a lot in the first three classes.

The proofs in class are meant to illustrate techniques that can be used to tackle many algorithmic and data related problems. You do not need to have these proofs or their conclusions memorized.

- Know and be comfortable applying: independence, linearity of expectation, linearity of variance, union bound, Markov's inequality, Chebyshev's inequality, basic probability calculations.
- Able to use techniques like: breaking random variables into sums of indicator variables and analyzing expected collisions (used in CATCHA analysis, two-level hashing analysis, *and* variance calculations for randomized load balancing.)
- Know definitions of 2-universal and pairwise independent hash functions and why they are useful.
- Able to apply exponential tail bounds (not have them memorized.)

A GENERAL COMMENT

We have covered a lot in the first three classes.

The proofs in class are meant to illustrate techniques that can be used to tackle many algorithmic and data related problems. You do not need to have these proofs or their conclusions memorized.

- Know and be comfortable applying: independence, linearity of expectation, linearity of variance, union bound, Markov's inequality, Chebyshev's inequality, basic probability calculations.
- Able to use techniques like: breaking random variables into sums of indicator variables and analyzing expected collisions (used in CATCHA analysis, two-level hashing analysis, *and* variance calculations for randomized load balancing.)
- Know definitions of 2-universal and pairwise independent hash functions and why they are useful.
- Able to apply exponential tail bounds (not have them memorized.)
- Understand law of large numbers and central limit theorem at a high level.

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$ and $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. For any $t \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

I messed up the math on these Bernstein inequality slides in class. Please refer to the non-annotated notes (posted on course site).

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$ and $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. For any $t \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Why can't this just applied when $n = 1$?

Bernstein Inequality: Consider ~~independent random~~ variables X_1~~, ..., X_n~~ in $[-M, M]$. Let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \text{Var}[X_1]$. For any $t \geq 0$:

$$\Pr\left(|X_1 - \mu| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Why can't this just applied when $n = 1$?

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}[X_1]$. For any $t \geq 0$:

$$\Pr\left(|X_1 - \mu| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Why can't this just applied when $n = 1$?

Observation 1: Only interesting to apply when $t \leq M$. Why?

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \text{Var}[X_1]$. For any $t \geq 0$:

$$\Pr\left(|X_1 - \mu| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Why can't this just applied when $n = 1$?

Observation 1: Only interesting to apply when $t \leq M$. Why?

Observation 2: When $t \leq M$,

$$2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right) \geq 2\exp(-3/4) \approx .95.$$

Where does sample size come in?

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$ and $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. For any $t \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Where does sample size come in?

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}[\sum_{i=1}^n X_i]$ and $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. For any $t \geq 0$:

$$\Pr\left(\left|\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + \frac{4}{3}Mt}\right)$$

Apply the inequality to the random variables $\frac{1}{n}X_1 \ldots, \frac{1}{n}X_n$.

Where does sample size come in?

Bernstein Inequality: Consider independent random variables $X_1, \ldots, X_n$ in $[-M, M]$. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]$ and $\sigma^2 = \text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$. For any $t \geq 0$:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{t^2 n}{2\sigma^2 + \frac{4}{3}Mt}\right).$$

Apply the inequality to the random variables $\frac{1}{n}X_1 \ldots, \frac{1}{n}X_n$.

Bound is $\leq \delta$ when $n \geq 2\log(1/\delta) \cdot \left(\frac{\sigma^2 + Mt}{t^2}\right)$

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

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. hash table.

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.

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

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

No false negatives.

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.

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

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

Solution: Bloom filters (repeated random hashing).

Chose k random hash functions $\mathsf{h}_1, \ldots, \mathsf{h}_k$ mapping the universe of elements $U \to [m]$.

Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.

Chose k random hash functions $\mathsf{h}_1, \ldots, \mathsf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathsf{h}_1(x)] = \ldots = A[\mathsf{h}_k(x)] := 1$.

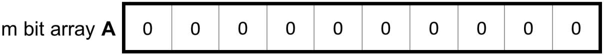
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.

BLOOM FILTERS

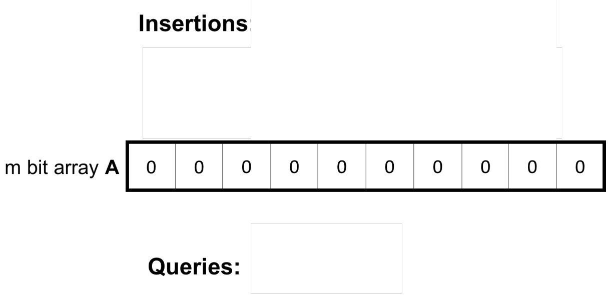
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



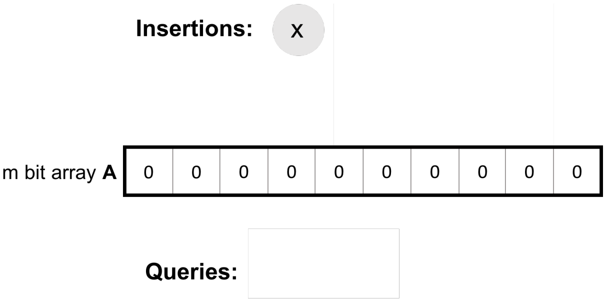
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



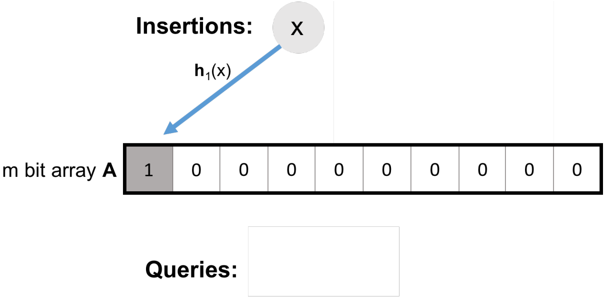
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



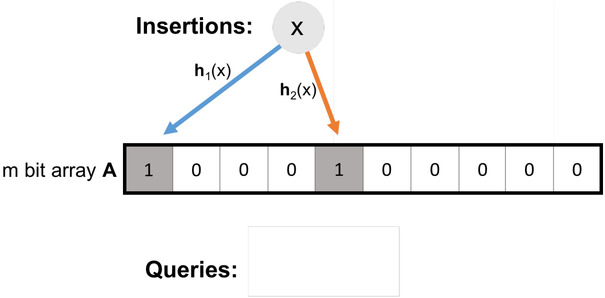
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



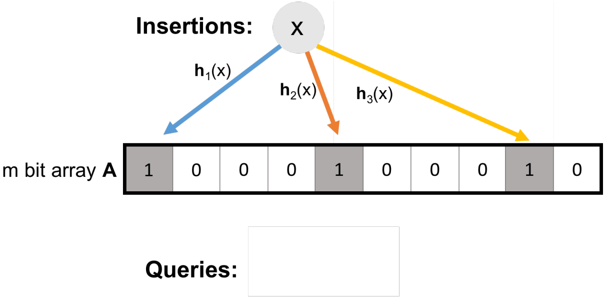
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- $insert(x)$: set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- $query(x)$: return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



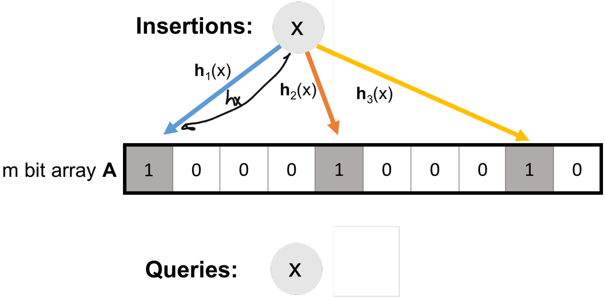
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



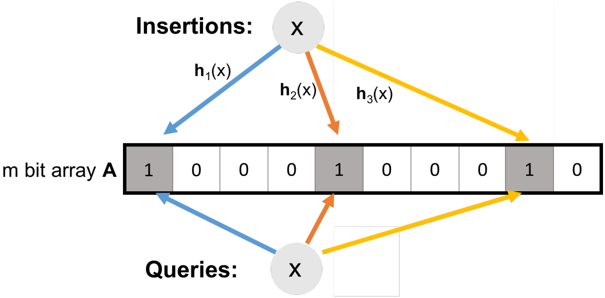
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



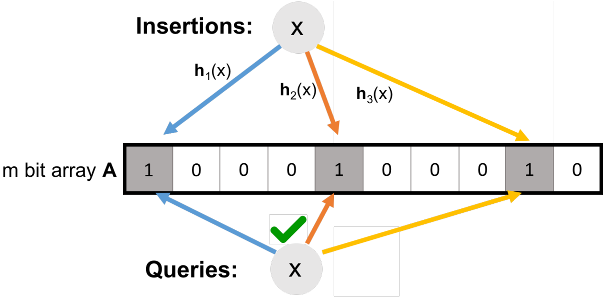
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- $insert(x)$: set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- $query(x)$: return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



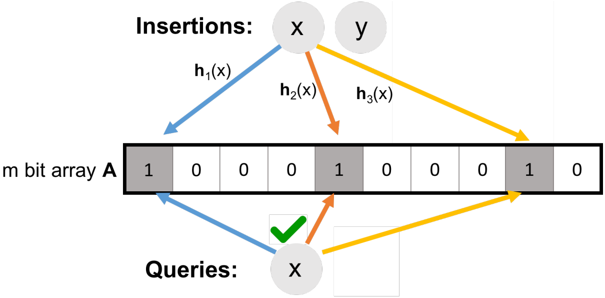
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



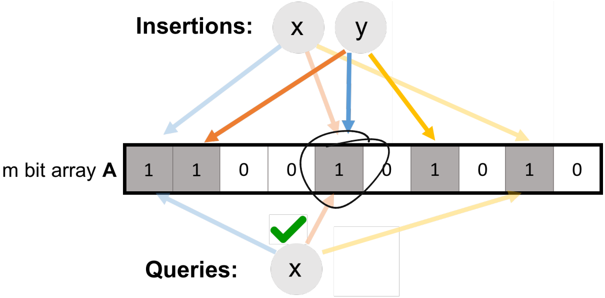
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



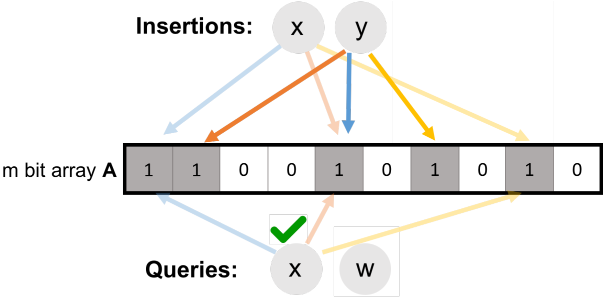
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



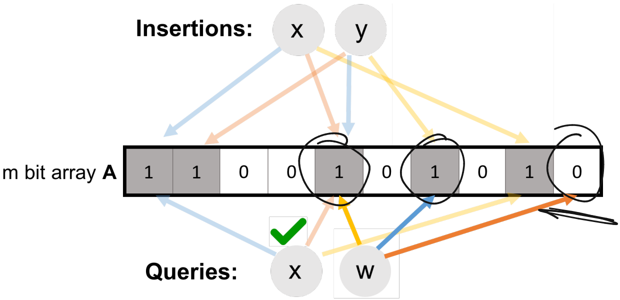
Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \rightarrow [m]$.

- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \to [m]$.

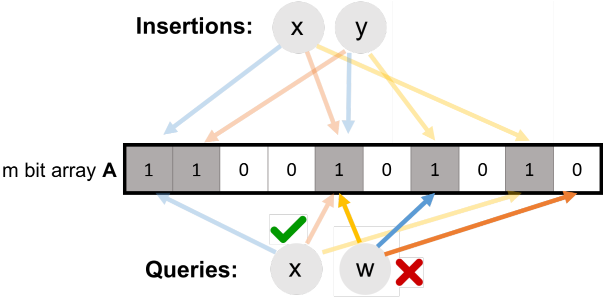
- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



BLOOM FILTERS

Chose k random hash functions $\mathbf{h}_1, \ldots, \mathbf{h}_k$ mapping the universe of elements $U \rightarrow [m]$.

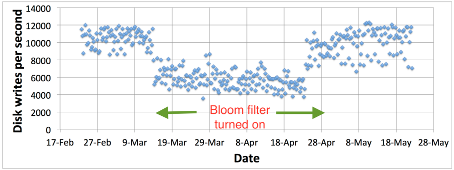
- Maintain an array A containing m bits, all initially 0.
- *insert*(x): set all bits $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] := 1$.
- *query*(x): return 1 only if $A[\mathbf{h}_1(x)] = \ldots = A[\mathbf{h}_k(x)] = 1$.



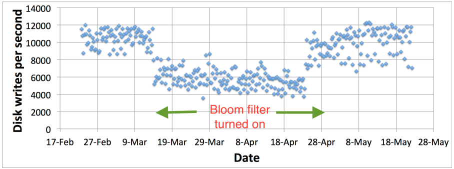
No false negatives. False positives more likely with more insertions.

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.



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.

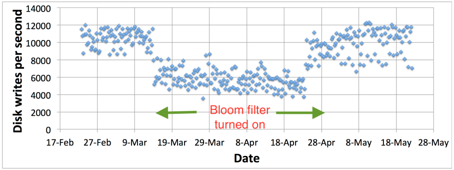


- When url x comes in, if $query(x) = 1$, cache the page at x. If not, run $insert(x)$ so that if it comes in again, it will be cached.

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.

Counting filters.
bloom

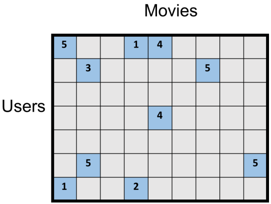


- When url x comes in, if $query(x) = 1$, cache the page at x. If not, run $insert(x)$ so that if it comes in again, it will be cached.
- **False positive:** A new url (possible one-hit-wonder) is cached. If the bloom filter has a false positive rate of $\delta = .05$, the number of cached one-hit-wonders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

APPLICATIONS: DATABASES

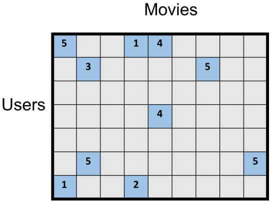
Distributed database systems, including Google Bigtable, Apache HBase, Apache Cassandra, and PostgreSQL use bloom filters to prevent expensive lookups of non-existent data.

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



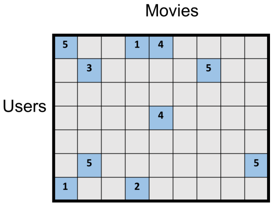
APPLICATIONS: DATABASES

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



- 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 requiring an out of memory access), check the bloom filter, which is stored in memory.

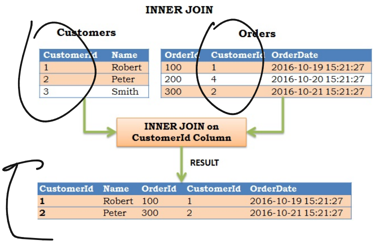
Distributed database systems, including Google Bigtable, Apache HBase, Apache Cassandra, and PostgreSQL use bloom filters to prevent expensive lookups of non-existent data.



- 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 requiring 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.

Bloom filters are used by Oracle and other database companies to speed up database *joins*.

Bloom filters are used by Oracle and other database companies to speed up database *joins*.



CustomerId	Name
1	Robert
2	Peter
3	Smith

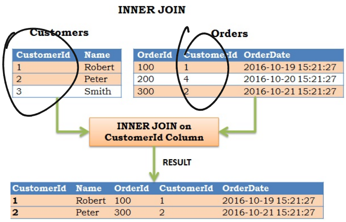
OrderId	CustomerId	OrderDate
100	1	2016-10-19 15:21:27
200	4	2016-10-20 15:21:27
300	2	2016-10-21 15:21:27

CustomerId	Name	OrderId	CustomerId	OrderDate
1	Robert	100	1	2016-10-19 15:21:27
2	Peter	300	2	2016-10-21 15:21:27

- Matches up a key in column **A** of one table to a key in column **B** of another, and merges corresponding information.

APPLICATIONS: DATABASES

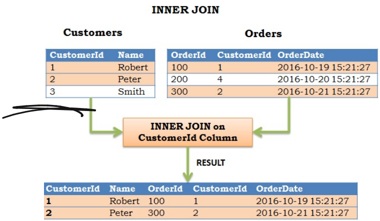
Bloom filters are used by Oracle and other database companies to speed up database *joins*.



- Matches up a key in column **A** of one table to a key in column **B** of another, and merges corresponding information.
- A bloom filter can be used to quickly eliminate entries that appear in **A** but not in **B**.

APPLICATIONS: DATABASES

Bloom filters are used by Oracle and other database companies to speed up database *joins*.



- Matches up a key in column **A** of one table to a key in column **B** of another, and merges corresponding information.
- A bloom filter can be used to quickly eliminate entries that appear in **A** but not in **B**.
- A false positive rate of δ means that a $1-\delta$ fraction of these entries can be eliminated in the initial bloom filter check.

- **Recommendation systems** (Netflix, Youtube, Tinder, etc.) use bloom filters to prevent showing users the same recommendations twice.

- **Recommendation systems** (Netflix, Youtube, Tinder, etc.) use bloom filters 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.

- Recommendation systems (Netflix, Youtube, Tinder, etc.) use bloom filters 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).

For a bloom filter with m bits and k hash functions, the insertion and query time is $O(k)$.

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 δ depend on m, k, and the number of items inserted?

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 δ depend on m, k, and the number of items inserted?

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

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 δ depend on m, k, and the number of items inserted?

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

$$\Pr(A[i] = 0) = \Pr\big(\mathbf{h}_1(x_1) \neq i \cap \ldots \cap \mathbf{h}_k(x_k) \neq i$$
$$\cap\, \mathbf{h}_1(x_2) \neq i \ldots \cap \mathbf{h}_k(x_2) \neq i \cap \ldots\big)$$

$n \cdot k$

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 δ depend on m, k, and the number of items inserted?

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

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

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 δ depend on m, k, and the number of items inserted?

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

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

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

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

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

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

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

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

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

$\frac{kn}{m}$

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

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

Step 1: What is the probability that after inserting n elements, the i^{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?

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

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

Step 1: What is the probability that after inserting n elements, the i^{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?

$$\begin{aligned}\Pr\left(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1\right) \\ = \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1)\end{aligned}$$

$1 - e^{-kn/m}$

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

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

Step 1: What is the probability that after inserting n elements, the i^{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?

$$\begin{aligned}\Pr\left(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1\right) \\ &= \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1) \\ &= \boxed{\left(1-e^{-\frac{kn}{m}}\right)^k}\end{aligned}$$

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

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

Step 1: What is the probability that after inserting n elements, the i^{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?

$$\begin{aligned}\Pr\left(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1\right) \\ &= \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1) \\ &= \left(1 - e^{-\frac{kn}{m}}\right)^k\end{aligned}$$

Actually Incorrect!

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

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

Step 1: What is the probability that after inserting n elements, the i^{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?

$$\begin{aligned}\Pr\left(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1\right) \\ &= \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1) \\ &= \left(1 - e^{-\frac{kn}{m}}\right)^k\end{aligned}$$

Actually Incorrect! Dependent events.

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

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:

$$\Pr(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1)$$
$$= \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1).$$

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

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:

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

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

CORRECT ANALYSIS SKETCH

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

$$\Pr(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1)$$
$$= \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1).$$

I.e., the events $A[\mathbf{h}_1(w)] = 1, \ldots, A[\mathbf{h}_k(w)] = 1$ 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, $\Pr(A[\mathbf{h}_j(w)] = 1) = \frac{t}{m}$.

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:

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

I.e., the events $A[\mathbf{h}_1(w)] = 1, \ldots, A[\mathbf{h}_k(w)] = 1$ 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, $\Pr(A[\mathbf{h}_j(w)] = 1) = \frac{t}{m}$.
- Thus conditioned on this event, the false positive rate is $\left(1 - \frac{t}{m}\right)^k$.

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:

$$\Pr(A[\mathbf{h}_1(w)] = \ldots = A[\mathbf{h}_k(w)] = 1)$$
$$= \Pr(A[\mathbf{h}_1(w)] = 1) \times \ldots \times \Pr(A[\mathbf{h}_k(w)] = 1).$$

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

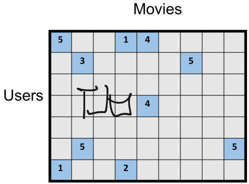
- Conditioned on this event, for any j, since $\mathbf{h}_j$ is a fully random hash function, $\Pr(A[\mathbf{h}_j(w)] = 1) = \frac{t}{m}$.
- Thus conditioned on this event, the false positive rate is $\left(1 - \frac{t}{m}\right)^k$.
- It remains to show that $\frac{t}{m} \approx e^{-\frac{kn}{m}}$ with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

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`

False Positive Rate: with m bits of storage, k hash functions, and n items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$.

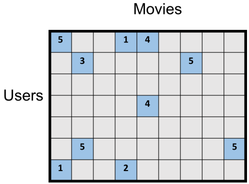
False Positive Rate: with m bits of storage, k hash functions, and n items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$.



- We have 100 million users and $10,000$ movies. On average each user has rated only 10 movies so of these 10^{12} possible (user,movie) pairs, only $10 * 100,000,000 = 10^9 = n$ (user,movie) pairs have non-empty entries in our table.
- We allocate $m = 8n = 8 \times 10^9$ bits for a Bloom filter (1 GB).

FALSE POSITIVE RATE

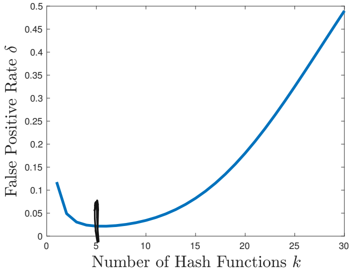
False Positive Rate: with m bits of storage, k hash functions, and n items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$.



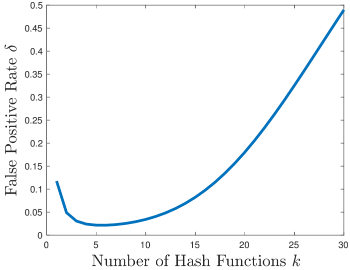
- We have 100 million users and $10,000$ movies. On average each user has rated only 10 movies so of these 10^{12} possible (user,movie) pairs, only $10 * 100,000,000 = 10^9 = n$ (user,movie) pairs have non-empty entries in our table.
- We allocate $m = 8n = 8 \times 10^9$ bits for a Bloom filter (1 GB). How should we set k to minimize the number of false positives?

False Positive Rate: with m bits of storage, k hash functions, and n items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$.

m = 8n
n = 10^9

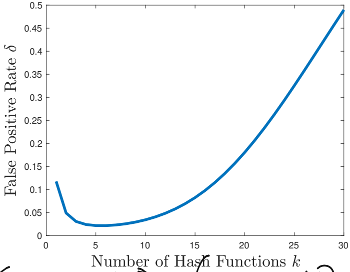


False Positive Rate: with m bits of storage, k hash functions, and n items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$.



- Can differentiate to show optimal number of hashes is $k = \ln 2 \cdot \frac{m}{n}$.

False Positive Rate: with m bits of storage, k hash functions, and n items inserted $\delta \approx \left(1 - e^{\frac{-kn}{m}}\right)^k$.



- Can differentiate to show optimal number of hashes is $k = \ln 2 \cdot \frac{m}{n}$.
- Balances between filling up the array with too many hashes and having enough hashes so that even when the array is pretty full, a new item is unlikely to have all its bits set (yield a false positive)

$m = 8n$ $1/8$ $k=1$ FP: $1/8$

$m = 8n$ $1/4$ $k=2$ FP: $\left(\frac{1}{4}\right)^2 = \frac{1}{16}$

What if we wanted to maintain a set with possible false negatives but no false positives?

SOMETHING TO THINK ABOUT

What if we wanted to maintain a set with possible false negatives but no false positives?

Turns out that this is extremely difficult.

Questions on Bloom Filters?

Stream Processing: Have a massive dataset X with n items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.

Stream Processing: Have a massive dataset X with n items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.
- Typically must compress the data on the fly, storing a data structure from which you can still learn useful information.

Stream Processing: Have a massive dataset X with n items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.
- Typically must compress the data on the fly, storing a data structure from which you can still learn useful information.
- Often the compression is randomized. E.g., bloom filters.

Stream Processing: Have a massive dataset X with n items $x_1, x_2, \ldots, x_n$ that arrive in a continuous stream. Not nearly enough space to store all the items (in a single location).

- Still want to analyze and learn from this data.
- Typically must compress the data on the fly, storing a data structure from which you can still learn useful information.
- Often the compression is randomized. E.g., bloom filters.
- Compared to traditional algorithm design, which focuses on minimizing runtime, the big question here is how much space is needed to answer queries of interest.

- **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.

- **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.

SOME EXAMPLES

- **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.
- **Internet Traffic**: 500 million Tweets per day, 5.6 billion Google searches, billions of ad-clicks and other logs from instrumented webpages, IPs routed by network switches, …

SOME EXAMPLES

- **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.
- **Internet Traffic**: 500 million Tweets per day, 5.6 billion Google searches, billions of ad-clicks and other logs from instrumented webpages, IPs routed by network switches, ...
- **Datasets in Machine Learning:** When training e.g. a neural network on a large dataset (ImageNet with 14 million images), the data is typically processed in a stream due to storage limitations.

SOME EXAMPLES

- **Sensor data:** images from telescopes (15 terabytes per night from the Large Synoptic Survey Telescope), readings from seismometer arrays monitoring and predicting earthquake activity, traffic cameras and travel time sensors (Smart Cities), electrical grid monitoring.
- **Internet Traffic**: 500 million Tweets per day, 5.6 billion Google searches, billions of ad-clicks and other logs from instrumented webpages, IPs routed by network switches, ...
- **Datasets in Machine Learning:** When training e.g. a neural network on a large dataset (ImageNet with 14 million images), the data is typically processed in a stream due to storage limitations.

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, output the number of distinct elements in the stream.

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, output the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$

Applications:

- Distinct IP addresses clicking on an ad or visiting a site.
- Distinct values in a database column (for estimating sizes of joins and group bys).
- Number of distinct search engine queries.
- Counting distinct motifs in large DNA sequences.

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements in the stream. E.g.,

$$1, 5, 7, 5, 2, 1 \rightarrow 4 \text{ distinct elements}$$

Applications:

- Distinct IP addresses clicking on an ad or visiting a site.
- Distinct values in a database column (for estimating sizes of joins and group bys).
- Number of distinct search engine queries.
- Counting distinct motifs in large DNA sequences.

Google Sawzall, Facebook Presto, Apache Drill, Twitter Algebird

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

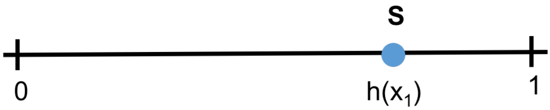
Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$

Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

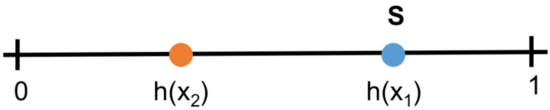
- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

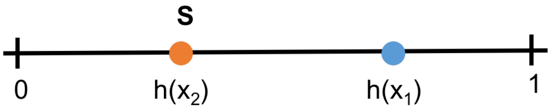
- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

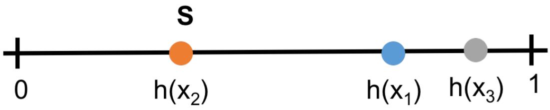
- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

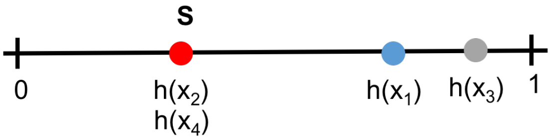
- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

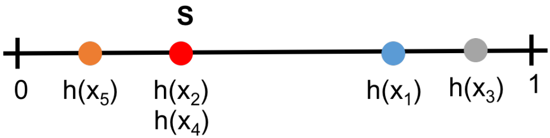
- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

Hashing for Distinct Elements (variant of Flajolet-Martin):

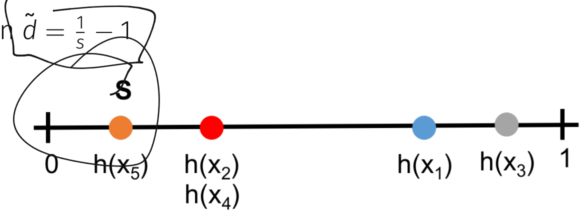
- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Distinct Elements (Count-Distinct) Problem: Given a stream $x_1, \ldots, x_n$, estimate the number of distinct elements.

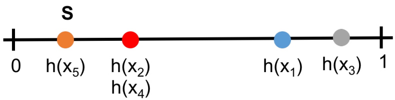
Hashing for Distinct Elements (variant of Flajolet-Martin):

- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



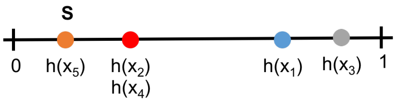
Hashing for Distinct Elements:

- Let $\mathbf{h} : U \rightarrow [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



Hashing for Distinct Elements:

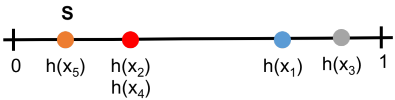
- Let $\mathbf{h} : U \to [0,1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



- After all items are processed, s is the minimum of d points chosen uniformly at random on $[0,1]$. Where $d = \#$ distinct elements.

Hashing for Distinct Elements:

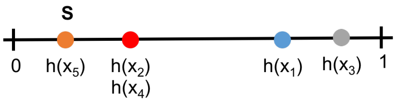
- Let $\mathbf{h} : U \to [0,1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$



- After all items are processed, s is the minimum of d points chosen uniformly at random on $[0,1]$. Where $d = \#$ distinct elements.
- Intuition: The larger d is, the smaller we expect s to be.

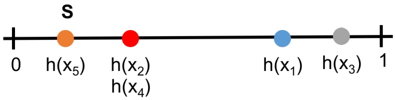
Hashing for Distinct Elements:

- Let $\mathbf{h} : U \to [0, 1]$ be a random hash function (with a real valued output)
- $s := 1$
- For $i = 1, \ldots, n$
  - $s := \min(s, \mathbf{h}(x_i))$
- Return $\tilde{d} = \frac{1}{s} - 1$

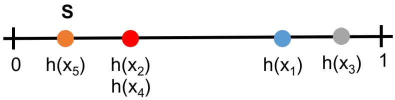


- After all items are processed, s is the minimum of d points chosen uniformly at random on $[0, 1]$. Where $d = \#$ distinct elements.
- Intuition: The larger d is, the smaller we expect s to be.
- Same idea as Flajolet-Martin algorithm and HyperLogLog, except they use discrete hash functions.

s computed by hashing algorithm is the minimum of d values chosen randomly in $[0, 1]$.

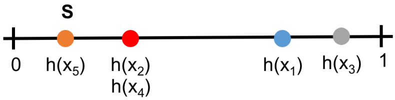


s computed by hashing algorithm is the minimum of d values chosen randomly in $[0, 1]$.



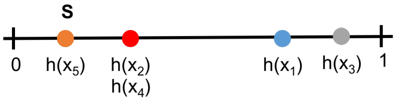
$$\mathbb{E}[s] =$$

s computed by hashing algorithm is the minimum of d values chosen randomly in $[0, 1]$.



$$\mathbb{E}[s] = \frac{1}{d+1}$$ (Interesting to prove to yourself.)

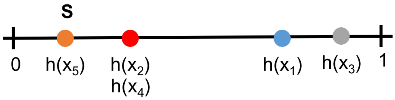
s computed by hashing algorithm is the minimum of d values chosen randomly in $[0, 1]$.



$$\mathbb{E}[s] = \frac{1}{d+1} \text{ (Interesting to prove to yourself.)}$$

- So estimate of $\tilde{d} = \frac{1}{s} - 1$ is correct if s exactly equals its expectation.

s computed by hashing algorithm is the minimum of d values chosen randomly in $[0, 1]$.



$$\mathbb{E}[s] = \frac{1}{d+1} \text{ (Interesting to prove to yourself.)}$$

- So estimate of $\tilde{d} = \frac{1}{s} - 1$ is correct if s exactly equals its expectation.
- If $|s - \mathbb{E}[s]| \leq \epsilon \cdot \mathbb{E}[s]$ for any $\epsilon \in (0, 1/2)$ can show:

$$(1 - 2\epsilon)d \leq \tilde{d} \leq (1 + 4\epsilon)d.$$

Next Time:

- Complete the analysis of hashing algorithm for distinct elements.
- Use a min-of-hashes technique for a different problem: estimating the similarity between two bit strings.

Next Time:

- Complete the analysis of hashing algorithm for distinct elements.
- Use a min-of-hashes technique for a different problem: estimating the similarity between two bit strings.

- A key idea behind audio fingerprint search (Shazam), document search (plagiarism and copyright violation detection), recommendation systems, etc.

Next Time:

- Complete the analysis of hashing algorithm for distinct elements.
- Use a min-of-hashes technique for a different problem: estimating the similarity between two bit strings.
- A key idea behind audio fingerprint search (Shazam), document search (plagiarism and copyright violation detection), recommendation systems, etc.

Questions?