Probabilistic Algorithms: What, Why, and How

A Deep Dive into Randomness in Computing

Sailesh Dahal

May 13, 2025

Kathmandu University

Outline

What are Probabilistic Algorithms?

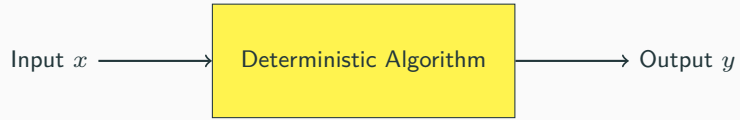
What is a Probabilistic Algorithm?

Definition
An algorithm that makes random choices during execution to influence its behavior or output.

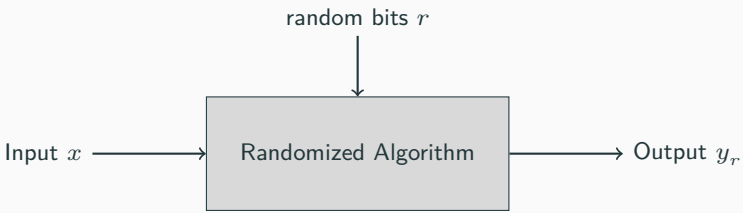
Deterministic vs Probabilistic Algorithm

Deterministic

Deterministic vs Probabilistic Algorithm



Deterministic



Probabilistic

Types of Probabilistic Algorithms

- Las Vegas Algorithms
- Monte Carlo Algorithms

Las Vegas Algorithms

Definition
Always produce a correct result, but the running time is a random variable (depends on random choices).

Las Vegas Algorithms

Definition
Always produce a correct result, but the running time is a random variable (depends on random choices).

- Output is always correct (no error), but time may vary between runs.

Las Vegas Algorithms

Definition
Always produce a correct result, but the running time is a random variable (depends on random choices).

- Output is always correct (no error), but time may vary between runs.
- Useful when correctness is critical, but we can tolerate variable performance.

Las Vegas Algorithms

Definition
Always produce a correct result, but the running time is a random variable (depends on random choices).

- Output is always correct (no error), but time may vary between runs.
- Useful when correctness is critical, but we can tolerate variable performance.
- **Example:** Randomized Quicksort
  - Randomly selects a pivot to avoid worst-case input.
  - Always sorts correctly, but running time depends on random choices.

Las Vegas Algorithms

Definition
Always produce a correct result, but the running time is a random variable (depends on random choices).

- Output is always correct (no error), but time may vary between runs.
- Useful when correctness is critical, but we can tolerate variable performance.
- **Example:** Randomized Quicksort
  - Randomly selects a pivot to avoid worst-case input.
  - Always sorts correctly, but running time depends on random choices.
- **Other Examples:**
  - Randomized algorithms for minimum cut in graphs (Karger's algorithm)

Las Vegas Algorithms

Definition
Always produce a correct result, but the running time is a random variable (depends on random choices).

- Output is always correct (no error), but time may vary between runs.
- Useful when correctness is critical, but we can tolerate variable performance.
- **Example:** Randomized Quicksort
  - Randomly selects a pivot to avoid worst-case input.
  - Always sorts correctly, but running time depends on random choices.
- **Other Examples:**
  - Randomized algorithms for minimum cut in graphs (Karger's algorithm)
- **Practical Usage in Software:**
  - Randomized Quicksort is widely used in standard libraries (e.g., C++ STL, Java, Python) for efficient sorting, especially to avoid worst-case performance on adversarial inputs.
  - Karger's algorithm is used in network reliability analysis tools.

Monte Carlo Algorithms

Definition
Always finish in a fixed amount of time, but may produce an incorrect result with a small probability.

Monte Carlo Algorithms

Definition
Always finish in a fixed amount of time, but may produce an incorrect result with a small probability.

- Running time is predictable (usually polynomial), but there is a chance of error.

Monte Carlo Algorithms

Definition
Always finish in a fixed amount of time, but may produce an incorrect result with a small probability.

- Running time is predictable (usually polynomial), but there is a chance of error.
- Useful when speed is more important than absolute certainty, and errors are rare or can be tolerated.

Monte Carlo Algorithms

Definition
Always finish in a fixed amount of time, but may produce an incorrect result with a small probability.

- Running time is predictable (usually polynomial), but there is a chance of error.
- Useful when speed is more important than absolute certainty, and errors are rare or can be tolerated.
- **Example:** Primality Testing (Miller-Rabin)
  - Quickly determines if a number is prime, but may err with small probability.
  - Error probability can be reduced by repeating the test.

Monte Carlo Algorithms

Definition
Always finish in a fixed amount of time, but may produce an incorrect result with a small probability.

- Running time is predictable (usually polynomial), but there is a chance of error.
- Useful when speed is more important than absolute certainty, and errors are rare or can be tolerated.
- **Example:** Primality Testing (Miller-Rabin)
  - Quickly determines if a number is prime, but may err with small probability.
  - Error probability can be reduced by repeating the test.
- **Other Examples:**
  - Monte Carlo integration (estimating π)
  - Randomized algorithms for approximate counting

Monte Carlo Algorithms

Definition
Always finish in a fixed amount of time, but may produce an incorrect result with a small probability.

- Running time is predictable (usually polynomial), but there is a chance of error.
- Useful when speed is more important than absolute certainty, and errors are rare or can be tolerated.
- **Example:** Primality Testing (Miller-Rabin)
  - Quickly determines if a number is prime, but may err with small probability.
  - Error probability can be reduced by repeating the test.
- **Other Examples:**
  - Monte Carlo integration (estimating π)
  - Randomized algorithms for approximate counting
- **Practical Usage in Software:**
  - Miller-Rabin primality test is used in cryptographic libraries (e.g., OpenSSL, GnuPG) for generating large prime numbers.

Applications of Probabilistic Algorithms

Real-World Motivation

- Web search (PageRank)
- Load balancing (power of two choices)
- Hashing (universal hash functions)
- Primality testing (Miller-Rabin)

Why Probabilistic Algorithms?

Why Randomness?

Motivation

- Simpler algorithms
- Better expected performance
- Avoid worst-case scenarios
- Useful for large-scale and distributed systems

How do Probabilistic Algorithms Work?

How: Randomization in Algorithms

Key Idea
Use random choices to influence the algorithm's path or output.

- Random pivot in Quicksort
- Random walks in graphs
- Random sampling

Example: Randomized Quicksort

QuickSort vs Randomized QuickSort

QuickSort:

1. Pick a pivot element from the array
2. Split array into 3 subarrays: those smaller than pivot, those larger than pivot, and the pivot itself
3. Recursively sort the subarrays, and concatenate them

Randomized QuickSort:

1. Pick a pivot element **uniformly at random** from the array
2. Split array into 3 subarrays: those smaller than pivot, those larger than pivot, and the pivot itself
3. Recursively sort the subarrays, and concatenate them

Example: Randomized Quicksort

Recall: QuickSort can take $\Omega(n^2)$ time to sort an array of size n.

Randomized QuickSort: Expected Runtime

Theorem

Randomized QuickSort sorts a given array of length n in $O(n \log n)$ expected time.

Note: On every input, randomized QuickSort takes $O(n \log n)$ time in expectation. On every input, it may take $\Omega(n^2)$ time with some small probability.

Randomized Quicksort: Step 1 (Initial Array)

Consider the array:

15	3	1	10	9	0	6	4

Randomized Quicksort: Step 1.1 (Pivot Chosen)

Suppose the random pivot chosen is 10 (at index 3):

Randomized Quicksort: Step 1.1 (Pivot Chosen)

Suppose the random pivot chosen is 10 (at index 3):

15	3	1	10	9	0	6	4

Randomized Quicksort: Step 2 (Partitioning Around Pivot 10)

After selecting pivot 10, we partition the array:

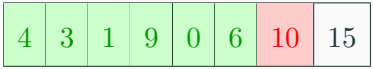
- Left: 4, 3, 1, 9, 0, 6 (elements before pivot position)
- Middle: 10 (pivot)
- Right: 15 (element after pivot position)

Randomized Quicksort: Step 2 (Partitioning Around Pivot 10)

After selecting pivot 10, we partition the array:

- Left: 4, 3, 1, 9, 0, 6 (elements before pivot position)
- Middle: 10 (pivot)
- Right: 15 (element after pivot position)

After partitioning:



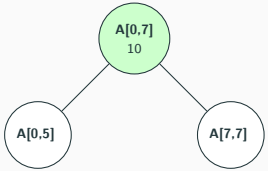
Randomized Quicksort: Step 2 (Partitioning Around Pivot 10)

After selecting pivot 10, we partition the array:

- Left: 4, 3, 1, 9, 0, 6 (elements before pivot position)
- Middle: 10 (pivot)
- Right: 15 (element after pivot position)

After partitioning:

4	3	1	9	0	6	10	15



Randomized Quicksort: Step 3 (Recurse Left [A[0,5]], Pivot 4)

Recurse on the left subarray:

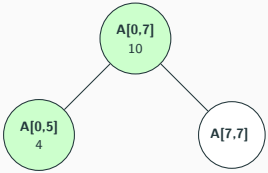
Let's choose a random pivot, say 4.

Randomized Quicksort: Step 3 (Recurse Left [A[0,5]], Pivot 4)

Recurse on the left subarray:

Let's choose a random pivot, say 4.

4	3	1	9	0	6



Randomized Quicksort: Step 3.1 (Partition Left [A[0,5]] Around 4)

After partitioning the left subarray:

Partition:

- Left: 0, 3, 1 (elements before pivot)
- Middle: 4 (pivot)
- Right: 9, 6 (elements after pivot)

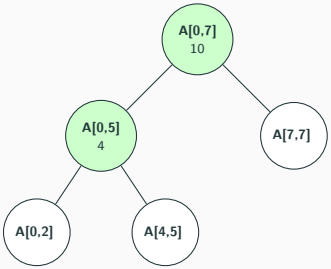
Randomized Quicksort: Step 3.1 (Partition Left [A[0,5]] Around 4)

After partitioning the left subarray:

0	3	1	4	9	6

Partition:

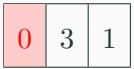
- Left: 0, 3, 1 (elements before pivot)
- Middle: 4 (pivot)
- Right: 9, 6 (elements after pivot)



Randomized Quicksort: Step 3.1.1 (Recurse Left [A[0,2]], Pivot 0)

Recurse on the left subarray:

Let's choose a random pivot, say 0.

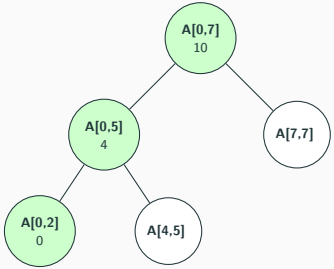


Randomized Quicksort: Step 3.1.1 (Recurse Left [A[0,2]], Pivot 0)

Recurse on the left subarray:

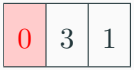
Let's choose a random pivot, say 0.

0	3	1



Randomized Quicksort: Step 3.1.1.1 (Partition Left [A[0,2]] Around 0)

After partitioning the left subarray:



Partition:

- Left: (empty)
- Middle: 0 (pivot)
- Right: 3, 1 (elements after pivot)

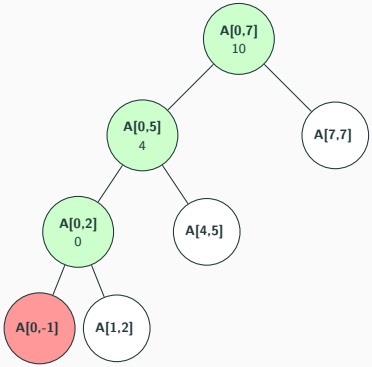
Randomized Quicksort: Step 3.1.1.1 (Partition Left [A[0,2]] Around 0)

After partitioning the left subarray:

0	3	1

Partition:

- Left: (empty)
- Middle: 0 (pivot)
- Right: 3, 1 (elements after pivot)



Randomized Quicksort: Step 3.1.1.2 (Recurse Right [A[1,2]], Pivot 3)

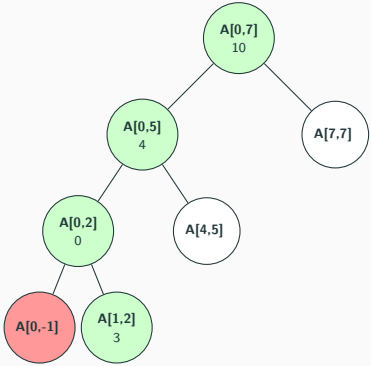
Recurse on the right subarray:

Let's choose a random pivot, say 3.

Randomized Quicksort: Step 3.1.1.2 (Recurse Right [A[1,2]], Pivot 3)

Recurse on the right subarray:

Let's choose a random pivot, say 3.



Randomized Quicksort: Step 3.1.1.2.1 (Partition [A[1,2]] Around 3)

After partitioning the left subarray:

Partition:

- Left: 1 (element before pivot)
- Middle: 3 (pivot)
- Right: (empty)

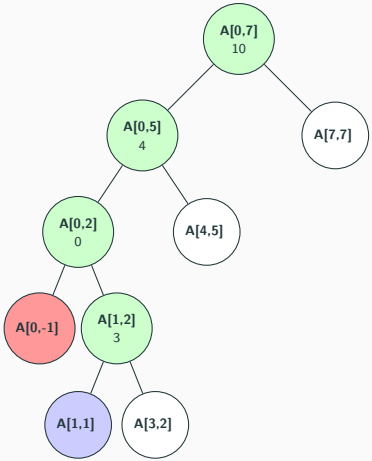
Randomized Quicksort: Step 3.1.1.2.1 (Partition [A[1,2]] Around 3)

After partitioning the left subarray:

1	3

Partition:

- Left: 1 (element before pivot)
- Middle: 3 (pivot)
- Right: (empty)



Randomized Quicksort: Step 3.1.1.2.1.1 (Recurse Left [A[1,1]], Done)

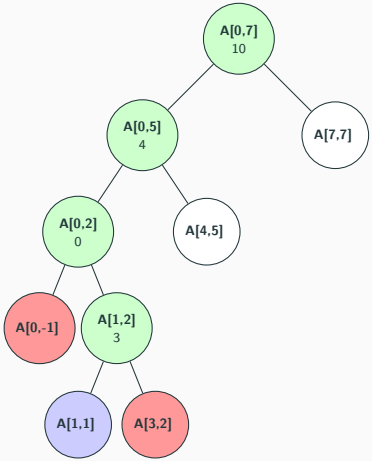
After partitioning the left subarray:

1

Partition:

- Left: (empty)
- Middle: 1 (pivot)
- Right: (empty)

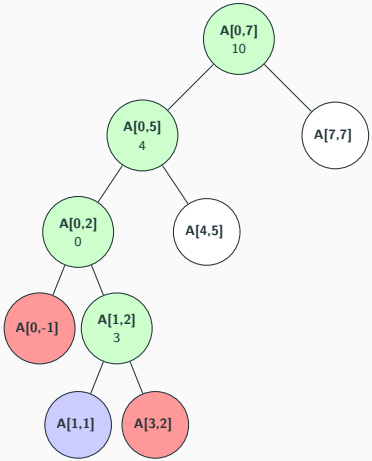
Single element subarray, done, return.



Randomized Quicksort: Step 3.1.1.2.1.2 (Recurse Right [A[3,2]], Done)

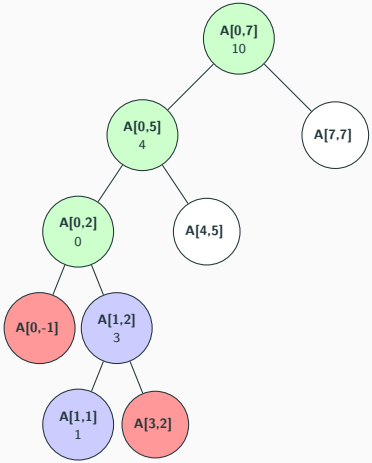
Recurse on the right subarray $A[3, 2]$ (empty, done).

Return to parent call $A[1, 2]$



Randomized Quicksort: Step 3.1.2 (Return to [A[0,2]])

Return to parent call $A[0, 2]$



Randomized Quicksort: Step 3.2 (Recurse Right [A[4,5]], Pivot 9)

Recurse on the right subarray $A[4, 5]$:

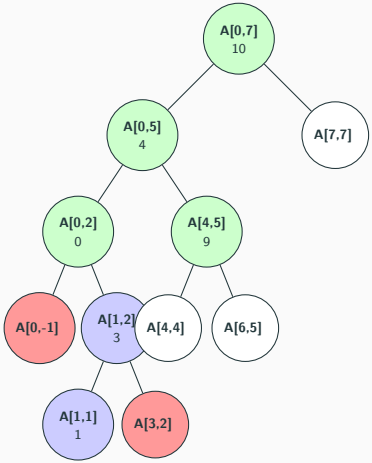
Let's choose a random pivot, say 9.

Randomized Quicksort: Step 3.2 (Recurse Right [A[4,5]], Pivot 9)

Recurse on the right subarray $A[4, 5]$:

Let's choose a random pivot, say 9.

9	6



Randomized Quicksort: Step 3.2.1 (Partition [A[4,5]] Around 9)

Recurse on the right subarray $A[4, 5]$:

Suppose the random pivot is 9:

Partition:

- Left: 6 (element before pivot)
- Middle: 9 (pivot)
- Right: (empty)

Randomized Quicksort: Step 3.2.1 (Partition [A[4,5]] Around 9)

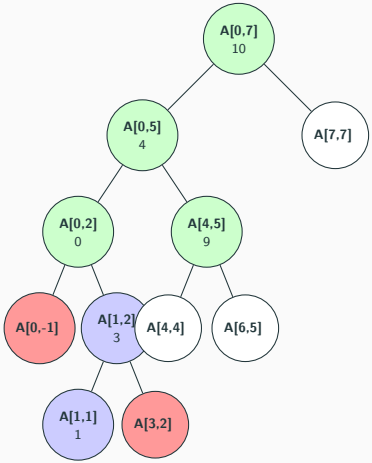
Recurse on the right subarray $A[4, 5]$:

Suppose the random pivot is 9:

6	9

Partition:

- Left: 6 (element before pivot)
- Middle: 9 (pivot)
- Right: (empty)



Randomized Quicksort: Step 3.2.1.1 (Recurse Left [A[4,4]], Done)

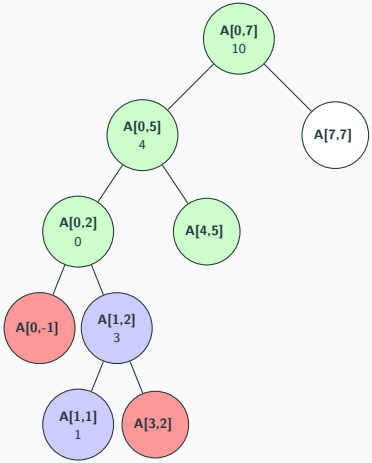
After partitioning the left subarray:

6

Partition:

- Left: (empty)
- Middle: 6 (pivot)
- Right: (empty)

Single element subarray, done, return.

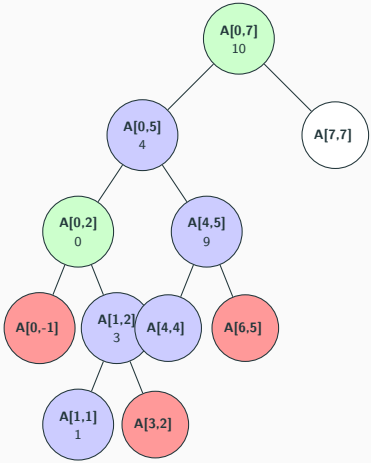


Randomized Quicksort: Step 3.2.1.2 (Recurse Right [A[6,5]], Done)

Recurse on the right subarray $A[6, 5]$ (empty, done).

Return to parent call $A[4, 5]$

Return to parent call $A[0, 5]$



Randomized Quicksort: Step 3.3 (Recurse Right [A[7,7]], Done)

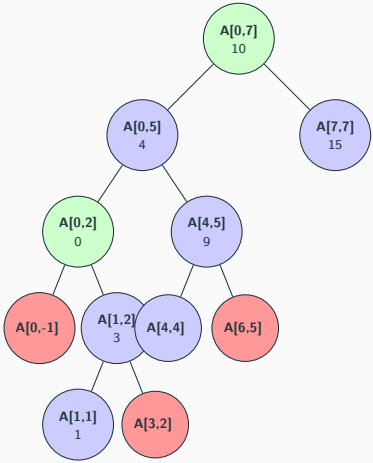
Recurse on the right subarray $A[7, 7]$ After partitioning the right subarray:

1

Partition:

- Left: (empty)
- Middle: 1 (pivot)
- Right: (empty)

Single element subarray, done, return.



Randomized Quicksort: Step 4 (Final Sorted Array)

The final sorted array is:

0	1	3	4	6	9	10	15

Quicksort Implementation (Python)

```python
import random
def quicksort(arr):
    if len(arr) <= 1:
        return arr
    pivot_idx = random.randrange(len(arr))  # Random index
    pivot = arr[pivot_idx]
    left = [x for x in arr if x < pivot]
    middle = [x for x in arr if x == pivot]
    right = [x for x in arr if x > pivot]
    return quicksort(left) + middle + quicksort(right)
```

Quicksort Recurrence

Expected Comparisons

$$T(n) \leq n + \frac{1}{n} \sum_{i=1}^{n} (T(i-1) + T(n-i))$$

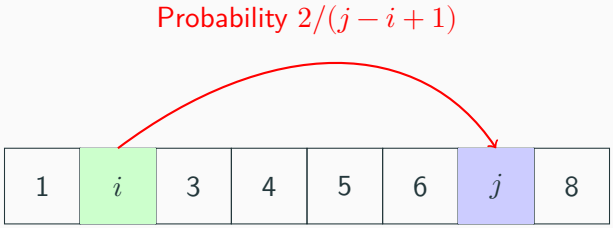
Base case: $T(1) = 0$

Solution: $T(n) = O(n \log n)$

Slick Analysis: Indicator Variables

- $Q(A)$: Number of comparisons on input A
- X_{ij}: Indicator for whether elements i and j are compared
- $E[Q(A)] = \sum_{i<j} Pr[R_{ij}]$
- $Pr[R_{ij}] = \frac{2}{j-i+1}$

Visualization



Harmonic Numbers in Analysis

Harmonic Number
$H_n = \sum_{i=1}^{n} \frac{1}{i} = \Theta(\log n)$

Summation in Quicksort
$E[Q(A)] \leq 2nH_n = O(n \log n)$

Example: Monte Carlo Estimation of pi

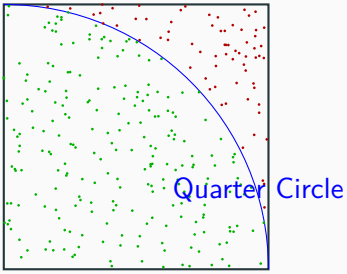
Monte Carlo Method - Estimating π

The Monte Carlo method estimates π by simulating random points in a unit square and counting how many fall inside a quarter circle of radius 1. The ratio of points inside the circle to the total points, multiplied by 4, approximates π.

Monte Carlo Algorithm

1. Generate N random points (x, y) where $0 \leq x \leq 1$ and $0 \leq y \leq 1$.
2. For each point, check if it lies inside the quarter circle: $x^2 + y^2 \leq 1$.
3. Count the number of points M that satisfy the condition.
4. Estimate π as: $\pi \approx 4 \times \frac{M}{N}$.

Visual Illustration



Example Calculation

- Suppose we generate $N = 1000$ random points in the unit square
- After simulation, we count $M = 785$ points inside the quarter circle
- We estimate π as:

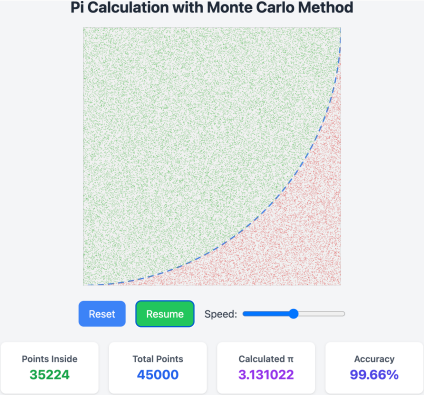
$$\pi \approx 4 \times \frac{M}{N} = 4 \times \frac{785}{1000} = 3.14$$

- The true value of π is approximately 3.14159

Convergence and Error Analysis

- The error in our estimate decreases as $O(1/\sqrt{N})$
- This means:
  - $N = 100$ points gives roughly 10% error
  - $N = 10,000$ points gives roughly 1% error
  - $N = 1,000,000$ points gives roughly 0.1% error
- The Monte Carlo method is especially useful for calculating multidimensional integrals
- For π calculation, there are more efficient methods, but this one is visually intuitive

Demo Visualization



Open Interactive Demo

Probabilistic Data Structures

Deterministic vs. Probabilistic Data Structures

Deterministic (e.g., Hash Set, List):

- Always provide exact answers.
- Can be space-intensive (store all elements).
- Operations might be slower for large datasets (e.g., disk I/O).
- **Guarantee:** No errors (false positives or negatives).

Probabilistic (e.g., Bloom Filter):

- Provide approximate answers with controlled error.
- Very space-efficient (use bits, not full elements).
- Operations are typically very fast (constant time).
- **Trade-off:** Small error probability for huge efficiency gains.

Key Idea
Use PDS when approximate answers are acceptable and space/speed are critical.

Example: Why PDS? Username Availability

The Problem
A website with millions of users needs to instantly check if a username is available during registration. How?

Deterministic Approach (Database Query):

- Store all usernames in a database.
- Query DB: ‘SELECT 1 FROM users WHERE username = ¿
- **Accurate? Yes.**
- **Fast? No.** Requires disk I/O, network latency.
- **Scalable? Poorly.** High load on DB servers.

Probabilistic Approach (Bloom Filter):

- Keep a compact Bloom filter in memory.
- Check filter: Is ‘username‘ possibly present?
- **Accurate? Mostly.** Small chance of false positive (saying taken when available), needs DB check then.
- **Fast? Yes.** In-memory check is O(k).
- **Scalable? Excellently.** Drastically reduces DB load.

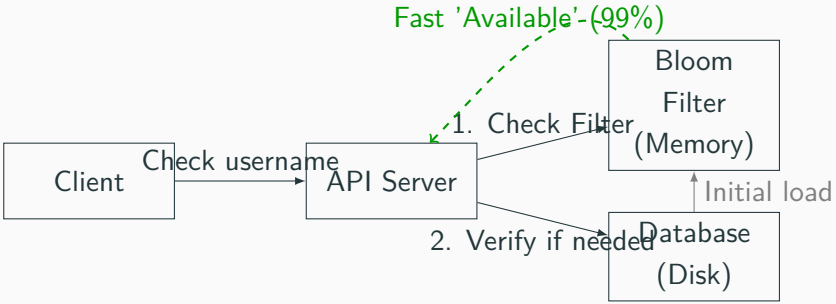
Username Checking: Implementation Details

1. **Initialization:** Load all existing usernames into Bloom filter at service startup (only infrequent DB reads).
2. **New registrations:** Add username to both database and Bloom filter.
3. **Availability check process:**
   - Check username against Bloom filter first (Fast, in-memory)
   - If Bloom filter says "definitely not in set" $\rightarrow$ Username is available (99% case for 1% error rate)
   - If Bloom filter says "possibly in set" $\rightarrow$ Verify with database query (Slow, but rare)

Performance Impact (10M users, 1% error)

- Memory: ≈ 18 MB Bloom Filter vs. hundreds of MB for DB index/cache.
- Speed: 99% of availability checks avoid slow database queries.

Username Checking: System Architecture



- Bloom filter acts as a fast, efficient preliminary check.
- Deterministic check (DB) used only as a fallback.
- Massively reduces load on the expensive resource (Database).

Bloom Filters: The Theory

- Space-efficient probabilistic data structure
- Tests if an element is a member of a set
- Possible false positives, never false negatives
- Components:
  - Bit array of m bits (initially all 0)
  - k independent hash functions

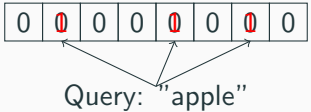
Bloom Filter Operations

Add element:

1. Hash element with k functions
2. Set bits at these k positions to 1

Query element:

1. Hash element with k functions
2. Check bits at these k positions
3. If **any** bit is 0: Definitely not in set
4. If **all** bits are 1: Probably in set



The Math Behind Bloom Filters

- **False positive probability (p):**

$$p \approx \left(1 - e^{-kn/m}\right)^k \tag{1}$$

- **Optimal size (m bits) for n items, error p:**

$$m = -\frac{n \ln p}{(\ln 2)^2} \tag{2}$$

- **Optimal hash functions (k):**

$$k = \frac{m}{n} \ln 2 \approx 0.7 \cdot \frac{m}{n} \tag{3}$$

Time and Space Complexity

Structure	Space	Lookup	Insert	Error Type
Hash Set	$O(n)$	$O(1)$ avg	$O(1)$ avg	None
Bloom Filter	$O(m)$	$O(k)$	$O(k)$	False Positives
Sorted List	$O(n)$	$O(\log n)$	$O(n)$	None
Trie	$O(N)$	$O(L)$	$O(L)$	None

n=items, m=bits ($m \ll n \times item\_size$), k=hashes, N=total chars, L=key length

Other Applications of Bloom Filters

Web/Database:

- Cache hit/miss optimization (e.g., CDNs)
- Avoid unnecessary DB lookups (like username example)
- Recommendation systems (seen items)

Other Applications of Bloom Filters

Web/Database:

- Cache hit/miss optimization (e.g., CDNs)
- Avoid unnecessary DB lookups (like username example)
- Recommendation systems (seen items)

Network:

- Web crawler URL deduplication (avoid re-crawling)
- Network packet routing (track flows efficiently)
- P2P network resource discovery

Other Applications of Bloom Filters

Web/Database:

- Cache hit/miss optimization (e.g., CDNs)
- Avoid unnecessary DB lookups (like username example)
- Recommendation systems (seen items)

Network:

- Web crawler URL deduplication (avoid re-crawling)
- Network packet routing (track flows efficiently)
- P2P network resource discovery

Security:

- Malware signature detection
- Spam filtering (known bad IPs/domains)
- Password breach checking (HaveIBeenPwned)

Other Applications of Bloom Filters

Web/Database:

- Cache hit/miss optimization (e.g., CDNs)
- Avoid unnecessary DB lookups (like username example)
- Recommendation systems (seen items)

Network:

- Web crawler URL deduplication (avoid re-crawling)
- Network packet routing (track flows efficiently)
- P2P network resource discovery

Security:

- Malware signature detection
- Spam filtering (known bad IPs/domains)
- Password breach checking (HaveIBeenPwned)

Big Data:

- Stream deduplication (unique visitors/events)
- Distributed data sync (approximate differences)
- Genomics (k-mer counting)

Other Applications of Bloom Filters

Web/Database:

- Cache hit/miss optimization (e.g., CDNs)
- Avoid unnecessary DB lookups (like username example)
- Recommendation systems (seen items)

Network:

- Web crawler URL deduplication (avoid re-crawling)
- Network packet routing (track flows efficiently)
- P2P network resource discovery

Security:

- Malware signature detection
- Spam filtering (known bad IPs/domains)
- Password breach checking (HaveIBeenPwned)

Big Data:

- Stream deduplication (unique visitors/events)
- Distributed data sync (approximate differences)
- Genomics (k-mer counting)

When to Use Bloom Filters

Bloom filters are ideal when:

- Memory is a critical constraint (Big Data, embedded systems)
- False positives are acceptable (can be handled by a secondary check)
- False negatives are unacceptable (must find all true positives)
- Elements are expensive to store or compare
- Lookup speed is crucial (real-time systems)
- Deletions are not needed (or use variants like Counting Bloom Filters)

References i