CSCE 658: Randomized Algorithms

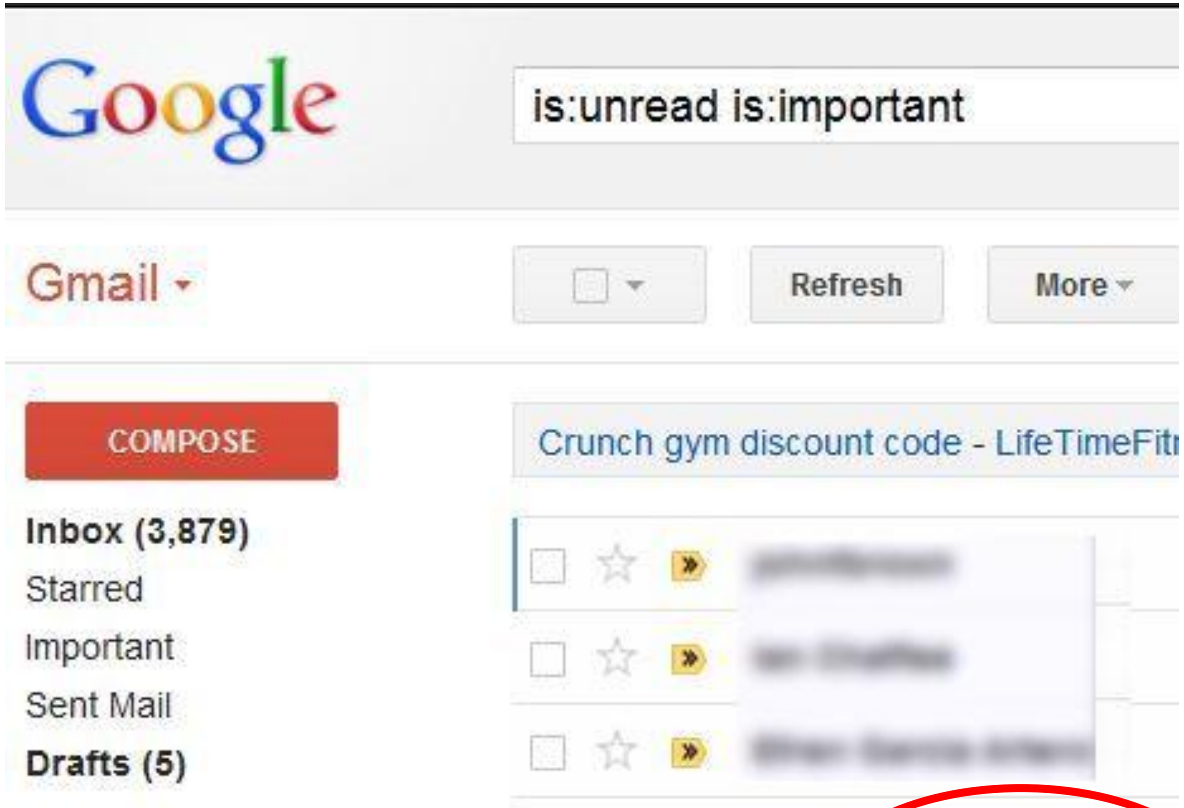
Lecture 1

Samson Zhou

Randomized Algorithms

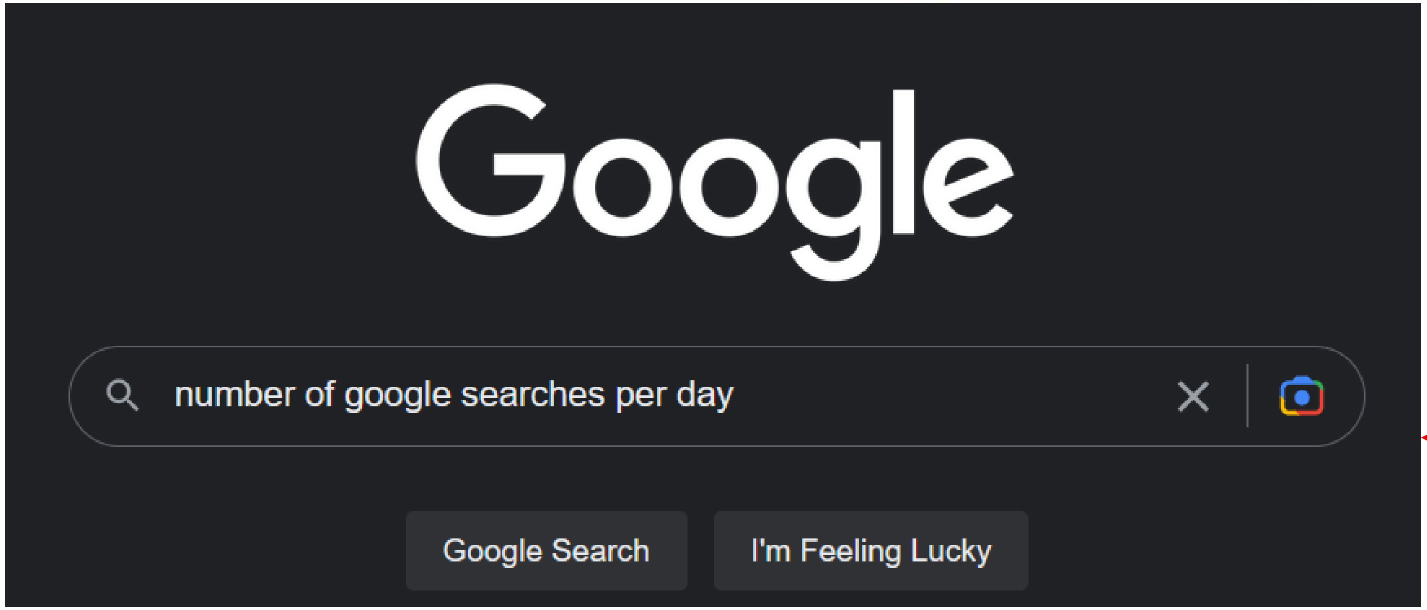
- A *randomized algorithm* is any algorithm that makes random choices during its execution, i.e., it uses randomly generated values to decide the each step of its execution
- The steps taken by a randomized algorithm might differ across multiple executions, even if the input remains the same
- The output may differ across multiple executions

Why Randomized Algorithms?



330 billion daily e-mails

3 billion monthly active users



8.5 billion daily Google searches

Randomized Algorithms

- Efficiency and Speed: Randomized algorithms can provide better runtimes compared to deterministic algorithms
  - Random sampling and random projects are used to handle large datasets

- Simplicity and Elegance: Randomized algorithms often offer solutions that are simpler and more elegant than deterministic counterparts
  - Primality testing can use randomization to efficiently determine whether a given number is like to be prime (Miller-Rabin primality testing)

Randomized Algorithms

- Probabilistic Guarantees: Errors may be acceptable in practice, e.g., input may be noisy or exact solutions are hard to achieve
  - Failure probabilities can often be tuned to only occur negligibly

- Avoiding Worst-Case Scenarios: Deterministic algorithms sometimes suffer from worst-case scenarios that might be unlikely to occur in practice. By introducing randomness, randomized algorithms can avoid being consistently unlucky and perform well on average.
  - Quicksort often performs better than deterministic counterparts, e.g., heapsort

2017 Equifax Data Breach

"Equifax agreed to a $700 million settlement over the privacy breach, but $425 million of that was set aside to repay consumers as a restitution fund."

YAHOO!

UBER

Gmail
by Google

Adobe

BLIZZARD
ENTERTAINMENT

eHarmony

AMERICA
Online

U.S. DEPARTMENT OF
HOMELAND SECURITY

LastPass ****

ebay

Dropbox

LinkedIn

PlayStation

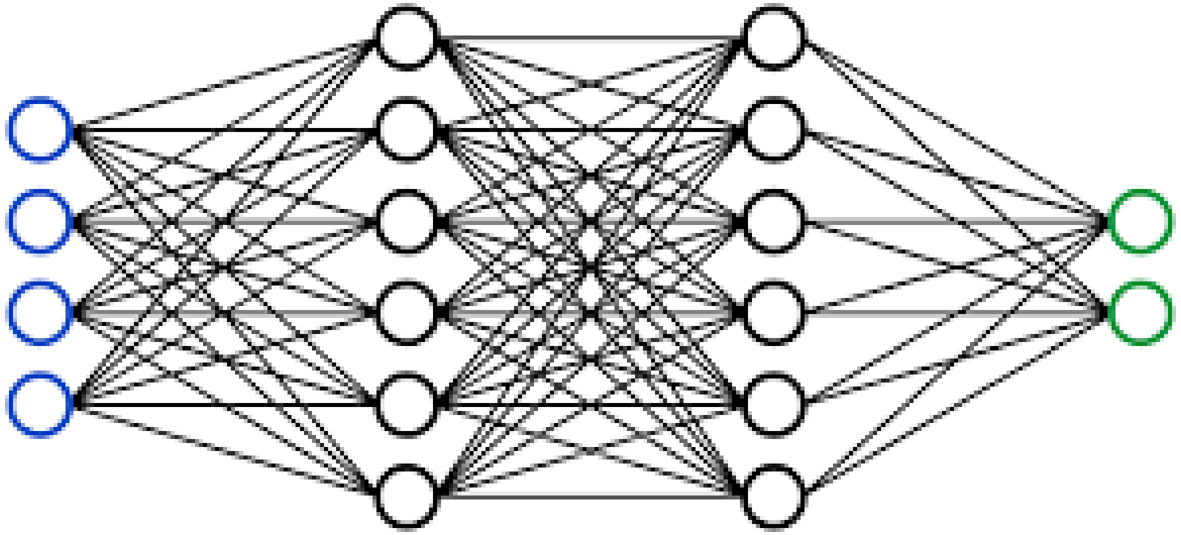
Randomized Algorithms

- Cryptography and Security: Randomized algorithms play a significant role in cryptography and security protocols.
  - They are used to generate random numbers, which is crucial for encryption, key generation, and other security-related processes.

- Privacy: Randomization is used to add noise to datasets to protect the privacy of individuals, while still maintaining accuracy.
  - Fundamental mechanisms to guarantee differential privacy (DP)

Randomized Algorithms

- Machine Learning and Data Analysis: Randomization is commonly used in machine learning
    - Randomization is used to prevent overfitting, augment data to increase diversity, perform mini-batch training, efficient tune hyperparameters



Toy Problem

- I am going to give you a string A and your partner a string B
- You are allowed to say a single digit (0-9) to your partner
- Your goal is to determine whether $A = B$ with probability 75%

Toy Problem

- $A = 7$

Toy Problem

- $B = 3$

Toy Problem

- $A = 7, B = 3$
- $A = 24$

Toy Problem

- $A = 7, B = 3$
- $B = 24$

Toy Problem

- $A = 7, B = 3$
- $A = 24, B = 24$
- $A = 7231$

Toy Problem

- $A = 7, B = 3$
- $A = 24, B = 24$
- $B = 7213$

Toy Problem

- $A = 7, B = 3$
- $A = 24, B = 24$
- $A = 7231, B = 7213$
- $A = 11112111$

Toy Problem

- $A = 7, B = 3$
- $A = 24, B = 24$
- $A = 7231, B = 7213$
- $B = 11111211$

Toy Problem

- $A = 7, B = 3$
- $A = 24, B = 24$
- $A = 7231, B = 7213$
- $A = 11112111, B = 11111211$

Logistics

- HRBB 126, TR, 5:30-6:45 pm CT
- Office Hours: PETR 424, 4:15-5:15 pm CT on Thursdays, or by appointment
- Course materials: https://samsonzhou.github.io/csce658-s24

Primary Goals

- Understand common tools for randomized algorithms
- Effectively and formally prove statements related to fundamental results in randomized algorithms, as measured by the homework problem sets
- Either:
  - Demonstrate the ability to conduct state-of-the-art research on randomized algorithms through a final project
  - Demonstrate the ability to design and analyze algorithms by leveraging the power of randomness, evaluated by a final examination

Secondary Goals

- Communicate technical ideas in a collaborative environment, as facilitated by the problem set groups (familiarity with LaTeX, practice communicating technical ideas)

Grading

- Group homework problem sets 50%
  - Groups of ≈5 students, one submission per group
  - Must be in LaTeX, submitted virtually via e-mail (or Canvas if necessary) by the deadline

- Final exam 50% OR final research project 50%
  - Group of n students: $4 + 6n$ page final report + final presentation
  - At least 10 research meetings with me over the semester

Useful Background

- Big Oh notation, e.g., $O(\log^{10} n)$, $O(\sqrt{n})$, $O(n^2)$
- Reductions, e.g., NP-hardness
- Mathematical maturity, exposure to reading and writing proofs

Questions?

Probability Basics

- Random variable (X)
- Sample space (Ω): Set of possible values (discrete/continuous, finite/infinite)
- Probability: $\Pr[X = x]$ represents the probability that the random variable X achieves value $x \in \Omega$

Joint and Conditional Probability

- Joint distribution: $\Pr[X = x, Y = y]$ is the probability X and Y achieve values x and y respectively

- Conditional distribution: $\Pr[X = x | Y = y]$ is the probability that X achieves the value x when Y achieves the value y

$$\Pr[X = x | Y = y] = \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}$$

- Marginal distribution: $\Pr[X = x] = \sum_{y \in \Omega_Y} \Pr[X = x | Y = y]$

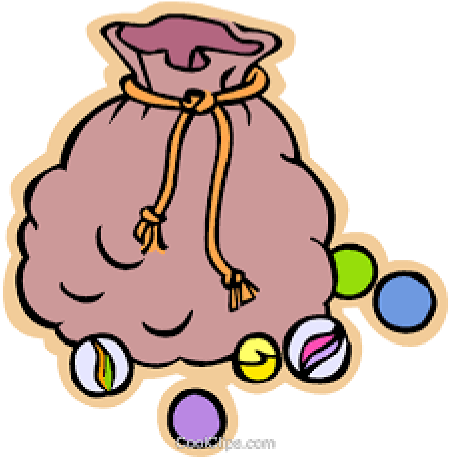
Independence

- Random variables X and Y are independent if $\Pr[X = x] = \Pr[X = x | Y = y]$ for all possible outcomes $x \in \Omega_X$, $y \in \Omega_Y$

Independence

- Suppose we have a bag with 1 red marble and 1 blue marble.
  - We draw a marble randomly from the bag
  - We put the marble back in the bag
  - We randomly draw another marble from the bag

- Let X be the color of the first marble drawn
- Let Y be the color of the second marble drawn



- Are X and Y independent?

Independence

- Suppose we have a bag with 1 red marble and 1 blue marble.
  - We draw a marble randomly from the bag
  - We DO NOT put the marble back in the bag
  - We randomly draw another marble from the bag

- Let X be the color of the first marble drawn
- Let Y be the color of the second marble drawn

- Are X and Y independent?

Boole's Inequality (Union Bound)

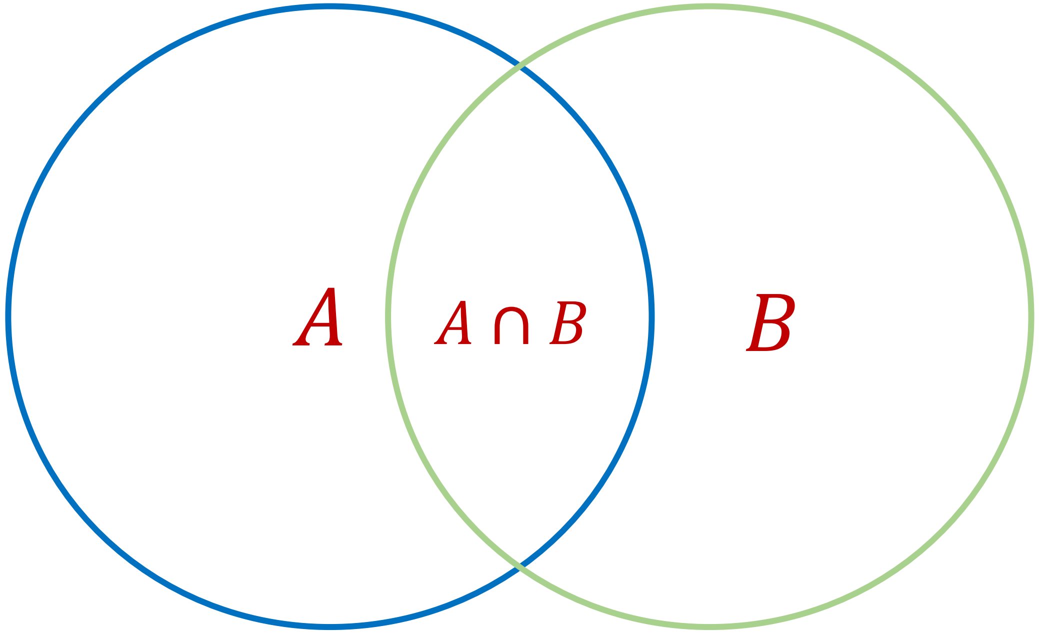
- Let $S_1, \ldots, S_k$ be a set of events that occur with probability $p_1, \ldots, p_k$

- The probability that at least one of the events $S_1, \ldots, S_k$ occurs is at most $p_1 + \cdots + p_k$

- Implication: the probability that NONE of the events $S_1, \ldots, S_k$ occur is at least $1 - (p_1 + \cdots + p_k)$

Boole's Inequality (Union Bound)

- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



- Proof by induction

Equality Problem

- Alice is given a string A and Bob is given a string B, each of length n, and they must determine whether $A = B$, using the *minimum amount of communication*

- Any deterministic protocol must use $\Omega(n)$ bits of communication, but there exists a randomized protocol that uses $O(\log n)$ bits of communication

Equality Problem

- Algorithm: Suppose Alice and Bob have access to a randomly generated string $x \in \{1,2,3,\ldots,q\}^n$. Alice sends over Ax and Bob determines whether $Ax = Bx$

- If $A = B$, then $Ax = Bx$ so the protocol succeeds
- If $A \neq B$, then what is the probability that $Ax \neq Bx$?

Schwartz-Zippel Lemma

- If $A \neq B$, then what is the probability that $Ax \neq Bx$, i.e., $(A - B)x \neq 0$?
- Note $(A - B)x$ is a linear polynomial in x

- [Schwartz-Zippel] Suppose P is a degree d polynomial in $x_1, \ldots, x_n$. Let $r_1, \ldots, r_n$ be randomly drawn from $\{1,2,3, \ldots, q\}$. Then
$$\Pr[P(r_1, \ldots, r_n) = 0] \leq \frac{d}{q}$$

Schwartz-Zippel Lemma

- [Schwartz-Zippel] Suppose P is a degree d polynomial in $x_1, \dots, x_n$. Let $r_1, \dots, r_n$ be randomly drawn from $\{1,2,3, \dots, q\}$. Then
$$\Pr[P(r_1, \dots, r_n) = 0] \leq \frac{d}{q}$$
- Proof by induction
- Base case: For $n = 1$, a degree d polynomial has d roots, so probability that r_1 hits a root is at most $\frac{d}{q}$
- Otherwise, write $P(x_1, \dots, x_n) = \sum_{i=0}^{d} x_1^i \cdot F_i(x_2, \dots, x_n)$

Schwartz-Zippel Lemma

- Since $P(x_1, \ldots, x_n) = \sum_{i=0}^{d} x_1^i \cdot F_i(x_2, \ldots, x_n)$ is nonzero, there exists nonzero $F_i(x_2, \ldots, x_n)$ with degree $d - i$
- Take the largest such i. By induction, $\Pr[F_i(r_2, \ldots, r_n) = 0] \leq \frac{d-i}{q}$
- Then $P(x_1, r_2 \ldots, r_n)$ is a polynomial of degree i so by induction, $\Pr[P(x_1, \ldots, x_n) = 0] \leq \frac{i}{q}$

Schwartz-Zippel Lemma

- Take the largest such i. By induction, $\Pr[F_i(r_2, \ldots, r_n) = 0] \leq \frac{d-i}{q}$
- Then $P(x_1, r_2 \ldots, r_n)$ is a polynomial of degree i so by induction, $\Pr[P(x_1, \ldots, x_n) = 0] \leq \frac{i}{q}$
- By union bound, $\Pr[P(r_1, \ldots, r_n) = 0] \leq \frac{d}{q}$

Equality Problem

- Algorithm: Suppose Alice and Bob have access to a randomly generated string $x \in \{1,2,3,\ldots,q\}^n$. Alice sends over Ax and Bob determines whether $Ax = Bx$

- If $A = B$, then $Ax = Bx$ so the protocol succeeds
- If $A \neq B$, then what is the probability that $Ax \neq Bx$?
- By Schwartz-Zippel, the probability that $Ax \neq Bx$ is at least $\frac{9}{10}$