CS 473: Fundamental Algorithms

Sariel Har-Peled, Alexandra Kolla
`sariel@illinois.edu`, `akolla@illinois.edu`,
3306 SC, 3222 SC

University of Illinois, Urbana-Champaign

Spring 2013

CS 473: Fundamental Algorithms, Spring 2013

Administrivia, Introduction, Graph basics and DFS

Lecture 1
January 15, 2013

The word “algorithm” comes from...

Muhammad ibn Musa al-Khwarizmi
780-850 AD
The word “algebra” is taken from the title of one of his books.

Part I

Administrivia

Instructional Staff

1. Instructor:
   - Sariel Har-Peled (`sariel`)
   - Alexandra Kolla (`akolla`)
2. Teaching Assistants:
   1. Danyal Khashabi (`khashab2`)
   2. Madan Vivek (`vmadan2`)
   3. Hai Wang (`hwang202`)
   4. Subhro Roy (`sroy9`)
3. Office hours: See course webpage
4. Email: See course webpage

Online resources

1. Webpage: `courses.engr.illinois.edu/cs473/sp2013/`
   General information, homeworks, etc.
2. Moodle:
   `https://learn.illinois.edu/course/view.php?id=1647`
   Quizzes, solutions to homeworks.
3. Online questions/announcements: Piazza
   `https://piazza.com/#spring2013/cs473`
   Online discussions, etc.

Textbooks

1. Prerequisites: CS 173 (discrete math), CS 225 (data structures) and CS 373 (theory of computation)
2. Recommended books:
   1. Algorithms by Dasgupta, Papadimitriou & Vazirani. Available online for free!
   2. Algorithm Design by Kleinberg & Tardos
3. Lecture notes: Available on the web-page after every class.
4. Additional References
   1. Previous class notes of Jeff Erickson, Sariel HarPeled and the instructor.
   2. Introduction to Algorithms: Cormen, Leiserson, Rivest, Stein.
   3. Computers and Intractability: Garey and Johnson.

Prerequisites

1. Asymptotic notation: $O(), \Omega(), o()$.
2. Discrete Structures: sets, functions, relations, equivalence classes, partial orders, trees, graphs
3. Logic: predicate logic, boolean algebra
4. Proofs: **by induction**, by contradiction
5. Basic sums and recurrences: sum of a geometric series, unrolling of recurrences, basic calculus
6. Data Structures: arrays, multi-dimensional arrays, linked lists, trees, balanced search trees, heaps
7. Abstract Data Types: lists, stacks, queues, dictionaries, priority queues
8. Algorithms: sorting (merge, quick, insertion), pre/post/in order traversal of trees, depth/breadth first search of trees (maybe graphs)
9. Basic analysis of algorithms: loops and nested loops, deriving recurrences from a recursive program
10. Concepts from Theory of Computation: languages, automata, Turing machine, undecidability, non-determinism
11. Programming: in some general purpose language
12. Elementary Discrete Probability: event, random variable, independence
13. Mathematical maturity

Grading Policy: Overview

1. Attendance/clickers: 5%
2. Quizzes: 5%
3. Homeworks: 20%
4. Midterms: 40% (**2 × 20%**)
5. Finals: 30% (covers the full course content)

Homeworks

1. One quiz every week: Due by midnight on Sunday.
2. One homework every week: Assigned on Tuesday and due the following Monday at noon.
3. Submit online only!
4. Homeworks can be worked on in groups of up to 3 and each group submits *one* written solution (except Homework 0).
   1. Short quiz-style questions to be answered individually on *Moodle*.
5. Groups can be changed a *few* times only
6. Unlike previous years no *oral* homework this semester due to large enrollment.

More on Homeworks

1. No extensions or late homeworks accepted.
2. To compensate, the homework with the least score will be dropped in calculating the homework average.
3. Important: Read homework faq/instructions on website.

Discussion Sessions

1. 50min problem solving session led by TAs
2. Four sections all in SC 1214.
   1. Tuesday
      5–5:50pm,
      6–6:50pm.
   2. Wednesday
      4–4:50pm,
      5–5:50pm.

Advice

1. Attend lectures, please ask plenty of questions.
2. Clickers...
3. Attend discussion sessions.
4. Don't skip homework and don't copy homework solutions.
5. Study regularly and keep up with the course.
6. Ask for help promptly. Make use of office hours.

Homeworks

1. HW 0 is posted on the class website. Quiz 0 available
2. Quiz 0 due by Sunday Jan 20 midnight
   HW 0 due on Monday January 21 noon.
3. Online submission.
4. HW 0 to be submitted in individually. f

Part II

Course Goals and Overview

Topics

1. Some fundamental algorithms
2. Broadly applicable techniques in algorithm design
   1. Understanding problem structure
   2. Brute force enumeration and backtrack search
   3. Reductions
   4. Recursion
      1. Divide and Conquer
      2. Dynamic Programming
   5. Greedy methods
   6. Network Flows and Linear/Integer Programming (optional)
3. Analysis techniques
   1. Correctness of algorithms via induction and other methods
   2. Recurrences
   3. Amortization and elementary potential functions
4. Polynomial-time Reductions, NP-Completeness, Heuristics

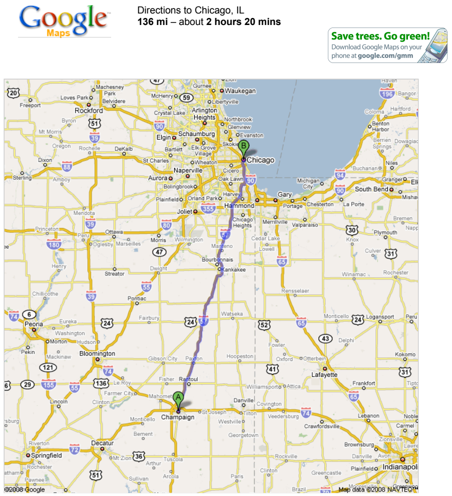
Goals

1.
2. Learn/remember some basic tricks, algorithms, problems, ideas
3. Understand/appreciate limits of computation (intractability)
4. Appreciate the importance of algorithms in computer science and beyond (engineering, mathematics, natural sciences, social sciences, ...)
5. Have fun!!!

Part III

Some Algorithmic Problems in the Real World

Shortest Paths



Directions to Chicago, IL
136 mi – about 2 hours 20 mins

Shortest Paths - Paris to Berlin

Übersichtskarte
Oslo
Uppsala
Stockholm
Göteborg
København
Leeds
Birmingham
Hamburg
Berlin
Lodz
War
Amsterdam
London
Bruxelles
Leipzig
Praha
Krakow
Paris
Luxembourg
Wien
Graz
Bern
200 km
Genève
(c)(c) AND, NAVTECH

Digital Information: Compression and Coding

Compression: reduce size for storage and transmission
Coding: add redundancy to protect against errors in storage and transmission

Efficient algorithms for compression/coding and decompressing/decoding part of most modern gadgets (computers, phones, music/video players ...)

Search and Indexing

String Matching and Link Analysis

1. Web search: Google, Yahoo!, Microsoft, Ask, ...
2. Text search: Text editors (Emacs, Word, Browsers, ...)
3. Regular expression search: grep, egrep, emacs, Perl, Awk, compilers

Public-Key Cryptography

Foundation of Electronic Commerce

RSA Crypto-system: generate key $\mathbf{n} = \mathbf{pq}$ where $\mathbf{p}, \mathbf{q}$ are *primes*

Primality: Given a number $\mathbf{N}$, check if $\mathbf{N}$ is a prime or composite.

Factoring: Given a composite number $\mathbf{N}$, find a non-trivial factor

Programming: Parsing and Debugging

[godavari: /temp/test] chekuri % gcc main.c

Parsing: Is main.c a syntactically valid C program?

Debugging: Will main.c go into an infinite loop on some input?

Easier problem ??? Will main.c halt on the specific input 10?

Optimization

Find the cheapest of most profitable way to do things

1. Airline schedules - AA, Delta, ...
2. Vehicle routing - trucking and transportation (UPS, FedEx, Union Pacific, ...)
3. Network Design - AT&T, Sprint, Level3 ...

Linear and Integer programming problems

Part IV

Algorithm Design

Important Ingredients in Algorithm Design

1. What is the problem (really)?
   1. What is the input? How is it represented?
   2. What is the output?
2. What is the model of computation? What basic operations are allowed?
3. Algorithm design
4. Analysis of correctness, running time, space etc.
5. Algorithmic engineering: evaluating and understanding of algorithm's performance in practice, performance tweaks, comparison with other algorithms etc. (Not covered in this course)

Primality testing

Problem

Given an integer $\mathbf{N} > \mathbf{0}$, is $\mathbf{N}$ a prime?

```
SimpleAlgorithm:
    for $i = 2$ to $\lfloor\sqrt{N}\rfloor$ do
        if $i$ divides $N$ then
            return ‘‘COMPOSITE’’
    return ‘‘PRIME’’
```

Correctness? If $\mathbf{N}$ is composite, at least one factor in $\{2, \ldots, \sqrt{\mathbf{N}}\}$

Running time? $\mathbf{O}(\sqrt{\mathbf{N}})$ divisions? Sub-linear in input size! Wrong!

Primality testing

...Polynomial means... in input size

How many bits to represent $\mathbf{N}$ in binary? $\lceil \log N \rceil$ bits.
Simple Algorithm takes $\sqrt{N} = 2^{(\log N)/2}$ time.
Exponential in the input size $\mathbf{n} = \log N$.

1. Modern cryptography: binary numbers with 128, 256, 512 bits.
2. Simple Algorithm will take 2^{64}, 2^{128}, 2^{256} steps!
3. Fastest computer today about 3 petaFlops/sec: 3×2^{50} floating point ops/sec.

Lesson:
Pay attention to representation size in analyzing efficiency of algorithms. Especially in *number* problems.

Efficient algorithms

So, is there an *efficient/good/effective* algorithm for primality?

Question:
What does efficiency mean?

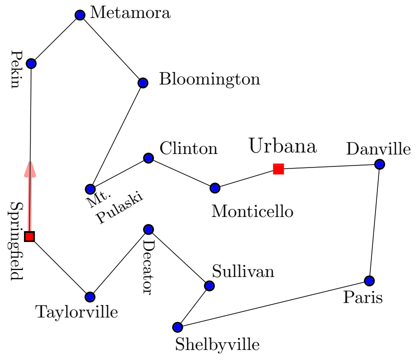
In this class *efficiency* is broadly equated to *polynomial time*.
$\mathbf{O(n), O(n \log n), O(n^2), O(n^3), O(n^{100}), \ldots}$ where $\mathbf{n}$ is size of the input.

Why? Is $\mathbf{n^{100}}$ really efficient/practical? Etc.

Short answer: polynomial time is a robust, mathematically sound way to define efficiency. Has been useful for several decades.

problem

Lincoln's tour



1. Circuit court - ride through counties staying a few days in each town.
2. Lincoln was a lawyer traveling with the Eighth Judicial Circuit.
3. Picture: travel during 1850.
   1. Very close to optimal tour.
   2. Might have been optimal at the time..

Solving by a Computer

Is it hard?

1. $\mathbf{n}$ = number of cities.
2. $\mathbf{n^2}$: size of input.
3. Number of possible solutions is
$$\mathbf{n * (n-1) * (n-2) * \ldots * 2 * 1 = n!}.$$
4. $\mathbf{n!}$ grows very quickly as $\mathbf{n}$ grows.
   $\mathbf{n = 10}$: $\mathbf{n! \approx 3628800}$
   $\mathbf{n = 50}$: $\mathbf{n! \approx 3 * 10^{64}}$
   $\mathbf{n = 100}$: $\mathbf{n! \approx 9 * 10^{157}}$

Solving by a Computer

Fastest computer...

1. Fastest super computer can do (roughly)

$$2.5 * 10^{15}$$

operations a second.

2. Assume: computer checks $2.5 * 10^{15}$ solutions every second, then...
   1. $n = 20 \implies$ 2 hours.
   2. $n = 25 \implies$ 200 years.
   3. $n = 37 \implies 2 * 10^{20}$ years!!!

What is a good algorithm?

Running time...

Input size	n^2 ops	n^3 ops	n^4 ops	$n!$ ops
5	0 secs	0 secs	0 secs	0 secs
20	0 secs	0 secs	0 secs	16 mins
30	0 secs	0 secs	0 secs	$3 \cdot 10^9$ years
100	0 secs	0 secs	0 secs	never
8000	0 secs	0 secs	1 secs	never
16000	0 secs	0 secs	26 secs	never
32000	0 secs	0 secs	6 mins	never
64000	0 secs	0 secs	111 mins	never
200,000	0 secs	3 secs	7 days	never
2,000,000	0 secs	53 mins	202.943 years	never
10^8	4 secs	12.6839 years	10^9 years	never
10^9	6 mins	12683.9 years	10^{13} years	never

What is a good algorithm?

Running time...

"No, Thursday's out. How about never—is never good for you?"

Primes is in P!

Theorem (Agrawal-Kayal-Saxena'02)

There is a polynomial time algorithm for primality.

First polynomial time algorithm for testing primality. Running time is $O(\log^{12} N)$ further improved to about $O(\log^6 N)$ by others. In terms of input size $n = \log N$, time is $O(n^6)$.

Breakthrough announced in August 2002. Three days later announced in New York Times. Only 9 pages!

Neeraj Kayal and Nitin Saxena were undergraduates at IIT-Kanpur!

What about before 2002?

Primality testing a key part of cryptography. What was the algorithm being used before 2002?

Miller-Rabin *randomized* algorithm:

1. runs in polynomial time: $\mathbf{O(\log^3 N)}$ time
2. if $\mathbf{N}$ is prime correctly says "yes".
3. if $\mathbf{N}$ is composite it says "yes" with probability at most $\mathbf{1/2^{100}}$ (can be reduced further at the expense of more running time).

Based on Fermat's little theorem and some basic number theory.

Factoring

1. Modern public-key cryptography based on RSA (Rivest-Shamir-Adelman) system.
2. Relies on the difficulty of factoring a composite number into its prime factors.
3. There is a polynomial time algorithm that decides whether a given number $\mathbf{N}$ is prime or not (hence composite or not) but no known polynomial time algorithm to factor a given number.

Lesson

Intractability can be useful!

Digression: decision, search and optimization

Three variants of problems.

1. Decision problem: answer is yes or no.
   Example: Given integer $\mathbf{N}$, is it a composite number?
2. Search problem: answer is a feasible solution if it exists.
   Example: Given integer $\mathbf{N}$, if $\mathbf{N}$ is composite output *a* non-trivial factor $\mathbf{p}$ of $\mathbf{N}$.
3. Optimization problem: answer is the *best* feasible solution (if one exists).
   Example: Given integer $\mathbf{N}$, if $\mathbf{N}$ is composite output the *smallest* non-trivial factor $\mathbf{p}$ of $\mathbf{N}$.

For a given underlying problem:

$$\text{Optimization} \geq \text{Search} \geq \text{Decision}$$

Quantum Computing

Theorem (Shor'1994)

There is a polynomial time algorithm for factoring on a quantum *computer.*

RSA and current commercial cryptographic systems can be broken if a quantum computer can be built!

Lesson

Pay attention to the model of computation.

Problems and Algorithms

Many many different problems.

1. Adding two numbers: efficient and simple algorithm
2. Sorting: efficient and not too difficult to design algorithm
3. Primality testing: simple and basic problem, took a long time to find efficient algorithm
4. Factoring: no efficient algorithm known.
5. Halting problem: important problem in practice, undecidable!

Multiplying Numbers

Problem Given two $\mathbf{n}$-digit numbers $\mathbf{x}$ and $\mathbf{y}$, compute their product.

Grade School Multiplication

Compute "partial product" by multiplying each digit of $\mathbf{y}$ with $\mathbf{x}$ and adding the partial products.

```
    3141
   ×2718
   25128
   3141
 21987
6282
 8537238
```

Time analysis of grade school multiplication

1. Each partial product: $\Theta(n)$ time
2. Number of partial products: $\leq n$
3. Adding partial products: $\mathbf{n}$ additions each $\Theta(n)$ (Why?)
4. Total time: $\Theta(n^2)$
5. Is there a faster way?

Fast Multiplication

Best known algorithm: $O(n \log n \cdot 2^{O(\log^* n)})$ time [Furer 2008]

Previous best time: $O(n \log n \log \log n)$ [Schonhage-Strassen 1971]

Conjecture: there exists and $O(n \log n)$ time algorithm

We don't fully understand multiplication!
Computation and algorithm design is non-trivial!

Course Approach

Algorithm design requires a mix of skill, experience, mathematical background/maturity and ingenuity.

Approach in this class and many others:

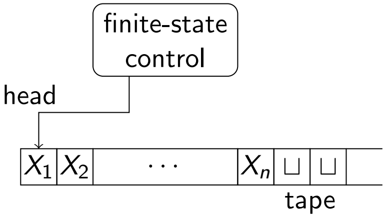
1. Improve skills by showing various tools in the abstract and with concrete examples
2. Improve experience by giving many problems to solve
3. Motivate and inspire
4. Creativity: you are on your own!

What model of computation do we use?

Turing Machine?

Turing Machines: Recap

1. Infinite tape
2. Finite state control
3. Input at beginning of tape
4. Special tape letter "blank" ⊔
5. Head can move only one cell to left or right



finite-state control

head

X_1 X_2 $\cdots$ X_n ⊔ ⊔

tape

Turing Machines

1. Basic unit of data is a bit (or a single character from a finite alphabet)
2. Algorithm is the finite control
3. Time is number of steps/head moves

Pros and Cons:

1. theoretically sound, robust and simple model that underpins computational complexity.
2. polynomial time equivalent to any reasonable "real" computer: Church-Turing thesis
3. too low-level and cumbersome, does not model actual computers for many realistic settings

"Real" Computers vs Turing Machines

How do "real" computers differ from TMs?

1. random access to memory
2. pointers
3. arithmetic operations (addition, subtraction, multiplication, division) in constant time

How do they do it?

1. basic data type is a word: currently 64 bits
2. arithmetic on words are basic instructions of computer
3. memory requirements assumed to be $\leq 2^{64}$ which allows for pointers and indirect addressing as well as random access

Unit-Cost RAM Model

Informal description:

1. Basic data type is an integer/floating point number
2. Numbers in input fit in a word
3. Arithmetic/comparison operations on words take constant time
4. Arrays allow random access (constant time to access $\mathbf{A[i]}$)
5. Pointer based data structures via storing addresses in a word

Example

Sorting: input is an array of $\mathbf{n}$ numbers

1. input size is $\mathbf{n}$ (ignore the bits in each number),
2. comparing two numbers takes $\mathbf{O(1)}$ time,
3. random access to array elements,
4. addition of indices takes constant time,
5. basic arithmetic operations take constant time,
6. reading/writing one word from/to memory takes constant time.

We will usually not allow (or be careful about allowing):

1. bitwise operations (and, or, xor, shift, etc).
2. floor function.
3. limit word size (usually assume unbounded word size).

Caveats of RAM Model

Unit-Cost RAM model is applicable in wide variety of settings in practice. However it is not a proper model in several important situations so one has to be careful.

1. For some problems such as basic arithmetic computation, unit-cost model makes no sense. Examples: multiplication of two $\mathbf{n}$-digit numbers, primality etc.
2. Input data is very large and does not satisfy the assumptions that individual numbers fit into a word or that total memory is bounded by $\mathbf{2^k}$ where $\mathbf{k}$ is word length.
3. Assumptions valid only for certain type of algorithms that do not create large numbers from initial data. For example, exponentiation creates very big numbers from initial numbers.

Models used in class

In this course:

1. Assume unit-cost RAM by default.
2. We will explicitly point out where unit-cost RAM is not applicable for the problem at hand.

Part V

Graph Basics

Why Graphs?

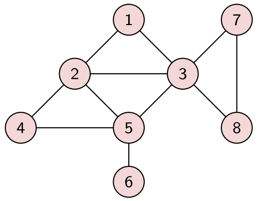
1. Graphs help model networks which are ubiquitous: transportation networks (rail, roads, airways), social networks (interpersonal relationships), information networks (web page links) etc etc.
2. Fundamental objects in Computer Science, Optimization, Combinatorics
3. Many important and useful optimization problems are graph problems
4. Graph theory: elegant, fun and deep mathematics

Graph

Definition

An undirected (simple) graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ is a $\mathbf{2}$-tuple:

1. $\mathbf{V}$ is a set of vertices (also referred to as nodes/points)
2. $\mathbf{E}$ is a set of edges where each edge $\mathbf{e} \in \mathbf{E}$ is a set of the form $\{\mathbf{u}, \mathbf{v}\}$ with $\mathbf{u}, \mathbf{v} \in \mathbf{V}$ and $\mathbf{u} \neq \mathbf{v}$.



Example

In figure, $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\mathbf{E} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{3, 5\}, \{3, 7\}, \{3, 8\}, \{4, 5\}, \{5, 6\}, \{7, 8\}\}$.

Notation and Convention

Notation

An edge in an undirected graphs is an *unordered* pair of nodes and hence it is a set. Conventionally we use $(\mathbf{u},\mathbf{v})$ for $\{\mathbf{u},\mathbf{v}\}$ when it is clear from the context that the graph is undirected.

1. $\mathbf{u}$ and $\mathbf{v}$ are the end points of an edge $\{\mathbf{u},\mathbf{v}\}$
2. Multi-graphs allow
    1. *loops* which are edges with the same node appearing as both end points
    2. *multi-edges*: different edges between same pairs of nodes
3. In this class we will assume that a graph is a simple graph unless explicitly stated otherwise.

Graph Representation I

Adjacency Matrix

Represent $\mathbf{G} = (\mathbf{V},\mathbf{E})$ with $\mathbf{n}$ vertices and $\mathbf{m}$ edges using a $\mathbf{n} \times \mathbf{n}$ adjacency matrix $\mathbf{A}$ where

1. $\mathbf{A[i,j] = A[j,i] = 1}$ if $\{\mathbf{i},\mathbf{j}\} \in \mathbf{E}$ and $\mathbf{A[i,j] = A[j,i] = 0}$ if $\{\mathbf{i},\mathbf{j}\} \not\in \mathbf{E}$.
2. Advantage: can check if $\{\mathbf{i},\mathbf{j}\} \in \mathbf{E}$ in $\mathbf{O(1)}$ time
3. Disadvantage: needs $\mathbf{\Omega(n^2)}$ space even when $\mathbf{m} \ll \mathbf{n^2}$

Graph Representation II

Adjacency Lists

Represent $\mathbf{G} = (\mathbf{V},\mathbf{E})$ with $\mathbf{n}$ vertices and $\mathbf{m}$ edges using adjacency lists:

1. For each $\mathbf{u} \in \mathbf{V}$, $\mathrm{Adj}(\mathbf{u}) = \{\mathbf{v} \mid \{\mathbf{u},\mathbf{v}\} \in \mathbf{E}\}$, that is neighbors of $\mathbf{u}$. Sometimes $\mathrm{Adj}(\mathbf{u})$ is the list of edges incident to $\mathbf{u}$.
2. Advantage: space is $\mathbf{O(m+n)}$
3. Disadvantage: cannot “easily” determine in $\mathbf{O(1)}$ time whether $\{\mathbf{i},\mathbf{j}\} \in \mathbf{E}$
    1. By sorting each list, one can achieve $\mathbf{O(\log n)}$ time
    2. By hashing “appropriately”, one can achieve $\mathbf{O(1)}$ time

Note: In this class we will assume that by default, graphs are represented using plain vanilla (unsorted) adjacency lists.

Connectivity

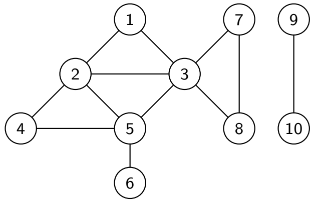
Given a graph $\mathbf{G} = (\mathbf{V},\mathbf{E})$:

1. A path is a sequence of *distinct* vertices $\mathbf{v_1, v_2, \ldots, v_k}$ such that $\{\mathbf{v_i, v_{i+1}}\} \in \mathbf{E}$ for $\mathbf{1 \leq i \leq k-1}$. The length of the path is $\mathbf{k-1}$ and the path is from $\mathbf{v_1}$ to $\mathbf{v_k}$
2. A cycle is a sequence of *distinct* vertices $\mathbf{v_1, v_2, \ldots, v_k}$ such that $\{\mathbf{v_i, v_{i+1}}\} \in \mathbf{E}$ for $\mathbf{1 \leq i \leq k-1}$ and $\{\mathbf{v_1, v_k}\} \in \mathbf{E}$.
3. A vertex $\mathbf{u}$ is connected to $\mathbf{v}$ if there is a path from $\mathbf{u}$ to $\mathbf{v}$.
4. The connected component of $\mathbf{u}$, $\mathrm{con}(\mathbf{u})$, is the set of all vertices connected to $\mathbf{u}$.

Connectivity contd

Define a relation $\mathbf{C}$ on $\mathbf{V} \times \mathbf{V}$ as $\mathbf{uCv}$ if $\mathbf{u}$ is connected to $\mathbf{v}$

1. In undirected graphs, connectivity is a reflexive, symmetric, and transitive relation. Connected components are the equivalence classes.
2. Graph is connected if only one connected component.



Connectivity Problems

Algorithmic Problems

1. Given graph $\mathbf{G}$ and nodes $\mathbf{u}$ and $\mathbf{v}$, is $\mathbf{u}$ *connected* to $\mathbf{v}$?
2. Given $\mathbf{G}$ and node $\mathbf{u}$, find all nodes that are connected to $\mathbf{u}$.
3. Find all connected components of $\mathbf{G}$.

Can be accomplished in $\mathbf{O(m+n)}$ time using **BFS** or **DFS**.

Basic Graph Search

Given $\mathbf{G = (V, E)}$ and vertex $\mathbf{u} \in \mathbf{V}$:

```
Explore(u):
    Initialize $S = \{u\}$
    while there is an edge $(x, y)$ with $x \in S$ and $y \notin S$ do
        add $y$ to $S$
```

Proposition

Explore$(\mathbf{u})$ *terminates with* $\mathbf{S} =$ *con*$(\mathbf{u})$.

Running time: depends on implementation

1. Breadth First Search (**BFS**): use queue data structure
2. Depth First Search (**DFS**): use stack data structure
3. Review CS 225 material!

Part VI

DFS

Depth First Search

DFS is a very versatile graph exploration strategy. Hopcroft and Tarjan (Turing Award winners) demonstrated the power of **DFS** to understand graph structure. **DFS** can be used to obtain linear time ($O(m + n)$) time algorithms for

1. Finding cut-edges and cut-vertices of undirected graphs
2. Finding strong connected components of directed graphs
3. Linear time algorithm for testing whether a graph is planar

DFS in Undirected Graphs

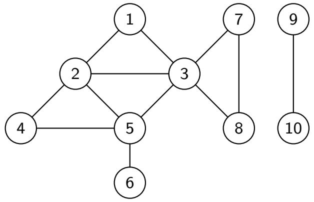
Recursive version.

```
DFS(G)
        Mark all nodes u as unvisited
        while there is an unvisited node u do
            DFS(u)

DFS(u)
        Mark u as visited
        for each edge (u,v) in Ajd(u) do
            if v is not marked
                DFS(v)
```

Implemented using a global array `Mark` for all recursive calls.

Example



Tree/Forest

```
DFS(G)
        Mark all nodes as unvisited
        T is set to ∅
        while ∃ unvisited node u do
            DFS(u)
        Output T
DFS(u)
        Mark u as visited
        for uv in Ajd(u) do
            if v is not marked
                add uv to T
                DFS(v)
```

Edges classified into two types: $uv \in E$ is a

1. tree edge: belongs to T
2. non-tree edge: does not belong to T

Properties of tree

Proposition

1. **T** *is a forest*
2. *connected components of* **T** *are same as those of* **G**.
3. *If* **uv** ∈ **E** *is a non-tree edge then, in* **T**, *either:*
   1. **u** *is an ancestor of* **v**, *or*
   2. **v** *is an ancestor of* **u**.

Question: Why are there no *cross-edges*?

with Visit Times

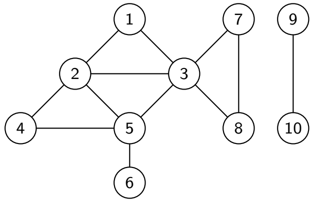
Keep track of when nodes are visited.

```
DFS(G)
    for all u ∈ V(G) do
        Mark u as unvisited
    T is set to ∅
    time = 0
    while ∃unvisited u do
        DFS(u)
    Output T
```

```
DFS(u)
    Mark u as visited
    pre(u) = ++time
    for each uv in Out(u) do
        if v is not marked then
            add edge uv to T
            DFS(v)
    post(u) = ++time
```

Scratch space

Example



pre and post numbers

Node $\mathbf{u}$ is **active** in time interval $[\mathrm{pre}(\mathbf{u}), \mathrm{post}(\mathbf{u})]$

Proposition

For any two nodes $\mathbf{u}$ *and* $\mathbf{v}$, *the two intervals* $[\mathrm{pre}(\mathbf{u}), \mathrm{post}(\mathbf{u})]$ *and* $[\mathrm{pre}(\mathbf{v}), \mathrm{post}(\mathbf{v})]$ *are disjoint or one is contained in the other.*

Proof.

- Assume without loss of generality that $\mathrm{pre}(\mathbf{u}) < \mathrm{pre}(\mathbf{v})$. Then $\mathbf{v}$ visited after $\mathbf{u}$.
- If **DFS**$(\mathbf{v})$ invoked before **DFS**$(\mathbf{u})$ finished, $\mathrm{post}(\mathbf{u}) > \mathrm{post}(\mathbf{v})$.
- If **DFS**$(\mathbf{v})$ invoked after **DFS**$(\mathbf{u})$ finished, $\mathrm{pre}(\mathbf{v}) > \mathrm{post}(\mathbf{u})$.

□

pre and post numbers useful in several applications of **DFS**- soon!

Part VII

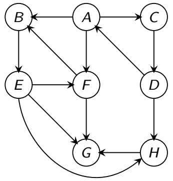
Directed Graphs and Decomposition

Directed Graphs

Definition

A directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ consists of

1. set of vertices/nodes $\mathbf{V}$ and
2. a set of edges/arcs $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$.



An edge is an *ordered* pair of vertices. $(\mathbf{u}, \mathbf{v})$ different from $(\mathbf{v}, \mathbf{u})$.

Examples of Directed Graphs

In many situations relationship between vertices is asymmetric:

1. Road networks with one-way streets.
2. Web-link graph: vertices are web-pages and there is an edge from page $\mathbf{p}$ to page $\mathbf{p'}$ if $\mathbf{p}$ has a link to $\mathbf{p'}$. Web graphs used by Google with PageRank algorithm to rank pages.
3. Dependency graphs in variety of applications: link from $\mathbf{x}$ to $\mathbf{y}$ if $\mathbf{y}$ depends on $\mathbf{x}$. Make files for compiling programs.
4. Program Analysis: functions/procedures are vertices and there is an edge from $\mathbf{x}$ to $\mathbf{y}$ if $\mathbf{x}$ calls $\mathbf{y}$.

Representation

Graph $\mathbf{G = (V, E)}$ with $\mathbf{n}$ vertices and $\mathbf{m}$ edges:

1. **Adjacency Matrix**: $\mathbf{n \times n}$ *asymmetric* matrix $\mathbf{A}$. $\mathbf{A[u, v] = 1}$ if $\mathbf{(u, v) \in E}$ and $\mathbf{A[u, v] = 0}$ if $\mathbf{(u, v) \not\in E}$. $\mathbf{A[u, v]}$ is not same as $\mathbf{A[v, u]}$.
2. **Adjacency Lists**: for each node $\mathbf{u}$, $\mathbf{Out(u)}$ (also referred to as $\mathbf{Adj(u)}$) and $\mathbf{In(u)}$ store out-going edges and in-coming edges from $\mathbf{u}$.

Default representation is adjacency lists.

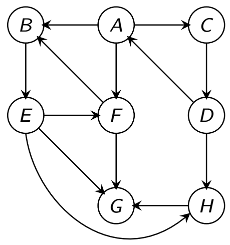
Directed Connectivity

Given a graph $\mathbf{G = (V, E)}$:

1. A **(directed) path** is a sequence of *distinct* vertices $\mathbf{v_1, v_2, \ldots, v_k}$ such that $\mathbf{(v_i, v_{i+1}) \in E}$ for $\mathbf{1 \leq i \leq k - 1}$. The length of the path is $\mathbf{k - 1}$ and the path is from $\mathbf{v_1}$ to $\mathbf{v_k}$
2. A **cycle** is a sequence of *distinct* vertices $\mathbf{v_1, v_2, \ldots, v_k}$ such that $\mathbf{(v_i, v_{i+1}) \in E}$ for $\mathbf{1 \leq i \leq k - 1}$ and $\mathbf{(v_k, v_1) \in E}$.
3. A vertex $\mathbf{u}$ can reach $\mathbf{v}$ if there is a path from $\mathbf{u}$ to $\mathbf{v}$. Alternatively $\mathbf{v}$ can be reached from $\mathbf{u}$
4. Let $\mathbf{rch(u)}$ be the set of all vertices reachable from $\mathbf{u}$.

Connectivity contd

Asymmetricity: $\mathbf{A}$ can reach $\mathbf{B}$ but $\mathbf{B}$ cannot reach $\mathbf{A}$



Questions:

1. Is there a notion of connected components?
2. How do we understand connectivity in directed graphs?

Connectivity and Strong Connected Components

Definition

Given a directed graph $\mathbf{G}$, $\mathbf{u}$ is strongly connected to $\mathbf{v}$ if $\mathbf{u}$ can reach $\mathbf{v}$ *and* $\mathbf{v}$ can reach $\mathbf{u}$. In other words $\mathbf{v} \in \mathrm{rch}\mathbf{(u)}$ and $\mathbf{u} \in \mathrm{rch}\mathbf{(v)}$.

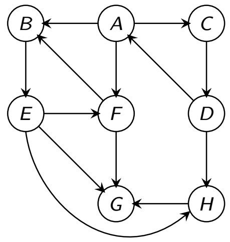
Define relation $\mathbf{C}$ where $\mathbf{uCv}$ if $\mathbf{u}$ is (strongly) connected to $\mathbf{v}$.

Proposition

$\mathbf{C}$ *is an equivalence relation, that is reflexive, symmetric and transitive.*

Equivalence classes of $\mathbf{C}$: *strong connected components* of $\mathbf{G}$.
They *partition* the vertices of $\mathbf{G}$.
$\mathrm{SCC}\mathbf{(u)}$: strongly connected component containing $\mathbf{u}$.

Strongly Connected Components: Example



Directed Graph Connectivity Problems

1. Given $\mathbf{G}$ and nodes $\mathbf{u}$ and $\mathbf{v}$, can $\mathbf{u}$ reach $\mathbf{v}$?
2. Given $\mathbf{G}$ and $\mathbf{u}$, compute rch$(\mathbf{u})$.
3. Given $\mathbf{G}$ and $\mathbf{u}$, compute all $\mathbf{v}$ that can reach $\mathbf{u}$, that is all $\mathbf{v}$ such that $\mathbf{u} \in$ rch$(\mathbf{v})$.
4. Find the strongly connected component containing node $\mathbf{u}$, that is $\mathrm{SCC}(\mathbf{u})$.
5. Is $\mathbf{G}$ strongly connected (a single strong component)?
6. Compute *all* strongly connected components of $\mathbf{G}$.

First four problems can be solve in $\mathbf{O(n+m)}$ time by adapting **BFS**/**DFS** to directed graphs. The last one requires a clever **DFS** based algorithm.

in Directed Graphs

```
DFS(G)
    Mark all nodes u as unvisited
    T is set to ∅
    time = 0
    while there is an unvisited node u do
        DFS(u)Output T

DFS(u)
    Mark u as visited
    pre(u) = ++time
    for each edge (u, v) in Out(u) do
        if v is not marked
            add edge (u, v) to T
            DFS(v)
    post(u) = ++time
```

DFS Properties

Generalizing ideas from undirected graphs:

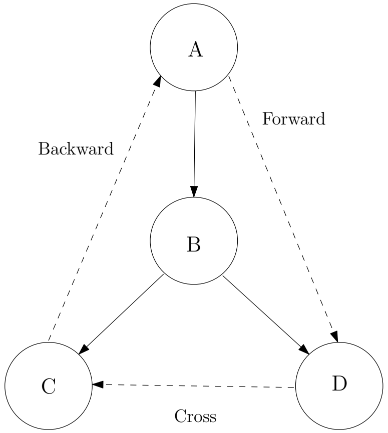
1. **DFS(u)** outputs a directed out-tree $\mathbf{T}$ rooted at $\mathbf{u}$
2. A vertex $\mathbf{v}$ is in $\mathbf{T}$ if and only if $\mathbf{v} \in$ rch$(\mathbf{u})$
3. For any two vertices $\mathbf{x}, \mathbf{y}$ the intervals $[\mathrm{pre}(\mathbf{x}), \mathrm{post}(\mathbf{x})]$ and $[\mathrm{pre}(\mathbf{y}), \mathrm{post}(\mathbf{y})]$ are either disjoint are one is contained in the other.
4. The running time of **DFS(u)** is $\mathbf{O(k)}$ where $\mathbf{k} = \sum_{\mathbf{v} \in \mathrm{rch}(\mathbf{u})} |\mathbf{Adj(v)}|$ plus the time to initialize the Mark array.
5. **DFS(G)** takes $\mathbf{O(m+n)}$ time. Edges in $\mathbf{T}$ form a disjoint collection of of out-trees. Output of **DFS(G)** depends on the order in which vertices are considered.

Tree

Edges of **G** can be classified with respect to the **DFS** tree **T** as:

1. **Tree edges** that belong to **T**
2. A **forward edge** is a non-tree edges (x, y) such that $\mathrm{pre}(x) < \mathrm{pre}(y) < \mathrm{post}(y) < \mathrm{post}(x)$.
3. A **backward edge** is a non-tree edge (x, y) such that $\mathrm{pre}(y) < \mathrm{pre}(x) < \mathrm{post}(x) < \mathrm{post}(y)$.
4. A **cross edge** is a non-tree edges (x, y) such that the intervals $[\mathrm{pre}(x), \mathrm{post}(x)]$ and $[\mathrm{pre}(y), \mathrm{post}(y)]$ are disjoint.

Types of Edges



Directed Graph Connectivity Problems

1. Given **G** and nodes **u** and **v**, can **u** reach **v**?
2. Given **G** and **u**, compute rch(u).
3. Given **G** and **u**, compute all **v** that can reach **u**, that is all **v** such that $u \in$ rch(v).
4. Find the strongly connected component containing node **u**, that is $\mathrm{SCC}(u)$.
5. Is **G** strongly connected (a single strong component)?
6. Compute *all* strongly connected components of **G**.

Algorithms via - I

1. Given **G** and nodes **u** and **v**, can **u** reach **v**?
2. Given **G** and **u**, compute rch(u).

Use **DFS(G, u)** to compute rch(u) in $O(n + m)$ time.

Algorithms via - II

1. Given $\mathbf{G}$ and $\mathbf{u}$, compute all $\mathbf{v}$ that can reach $\mathbf{u}$, that is all $\mathbf{v}$ such that $\mathbf{u} \in \mathrm{rch}(\mathbf{v})$.

Definition (Reverse graph.)

Given $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, $\mathbf{G}^{\mathrm{rev}}$ is the graph with edge directions reversed $\mathbf{G}^{\mathrm{rev}} = (\mathbf{V}, \mathbf{E}')$ where $\mathbf{E}' = \{(\mathbf{y}, \mathbf{x}) \mid (\mathbf{x}, \mathbf{y}) \in \mathbf{E}\}$

Compute $\mathrm{rch}(\mathbf{u})$ in $\mathbf{G}^{\mathrm{rev}}$!

1. **Correctness:** exercise
2. **Running time:** $\mathbf{O(n + m)}$ to obtain $\mathbf{G}^{\mathrm{rev}}$ from $\mathbf{G}$ and $\mathbf{O(n + m)}$ time to compute $\mathrm{rch}(\mathbf{u})$ via **DFS**. If both $\mathbf{Out(v)}$ and $\mathbf{In(v)}$ are available at each $\mathbf{v}$ then no need to explicitly compute $\mathbf{G}^{\mathrm{rev}}$. Can do it $\mathbf{DFS(u)}$ in $\mathbf{G}^{\mathrm{rev}}$ implicitly.

Algorithms via - III

$\mathbf{SC(G, u)} = \{\mathbf{v} \mid \mathbf{u}$ is strongly connected to $\mathbf{v}\}$

1. Find the strongly connected component containing node $\mathbf{u}$. That is, compute $\mathrm{SCC}(\mathbf{G}, \mathbf{u})$.

$$\mathrm{SCC}(\mathbf{G}, \mathbf{u}) = \mathrm{rch}(\mathbf{G}, \mathbf{u}) \cap \mathrm{rch}(\mathbf{G}^{\mathrm{rev}}, \mathbf{u})$$

Hence, $\mathrm{SCC}(\mathbf{G}, \mathbf{u})$ can be computed with two **DFS**es, one in $\mathbf{G}$ and the other in $\mathbf{G}^{\mathrm{rev}}$. Total $\mathbf{O(n + m)}$ time.

Algorithms via - IV

1. Is $\mathbf{G}$ strongly connected?

Pick arbitrary vertex $\mathbf{u}$. Check if $\mathbf{SC(G, u)} = \mathbf{V}$.

Algorithms via - V

1. Find *all* strongly connected components of $\mathbf{G}$.

```
for each vertex $\mathbf{u} \in \mathbf{V}$ do
        find $\mathbf{SC(G, u)}$
```

Running time: $\mathbf{O(n(n + m))}$.

Q: Can we do it in $\mathbf{O(n + m)}$ time?

Reading and Homework 0

Chapters 1 from Dasgupta etal book, Chapters 1-3 from Kleinberg-Tardos book.

Proving algorithms correct - Jeff Erickson's notes (see link on website)