COS 226, Fall 2013

Algorithms and Data Structures

PRINCETON UNIVERSITY

http://www.princeton.edu/~cos226

COS 226 course overview

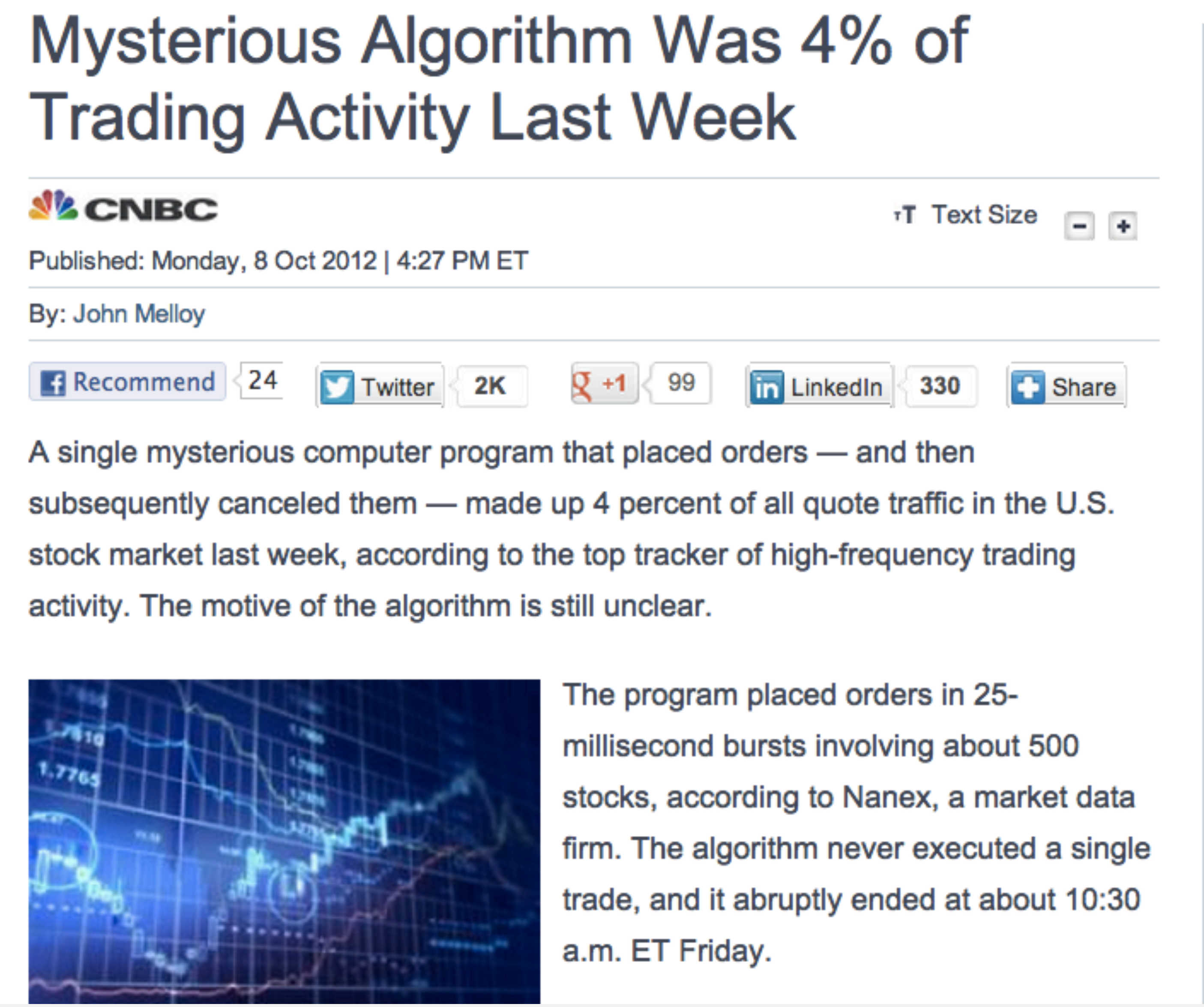
What is COS 226?

- Intermediate-level survey course.
- Programming and problem solving, with applications.
- Algorithm: method for solving a problem.
- Data structure: method to store information.

topic	data structures and algorithms
data types	stack, queue, bag, union-find, priority queue
sorting	quicksort, mergesort, heapsort, radix sorts
searching	BST, red-black BST, hash table
graphs	BFS, DFS, Prim, Kruskal, Dijkstra
strings	KMP, regular expressions, tries, data compression
advanced	B-tree, suffix array, maxflow, simplex

Why study algorithms?

Their impact is broad and far-reaching.



Mysterious Algorithm Was 4% of Trading Activity Last Week

Published: Monday, 8 Oct 2012 | 4:27 PM ET

By: John Melloy

A single mysterious computer program that placed orders — and then subsequently canceled them — made up 4 percent of all quote traffic in the U.S. stock market last week, according to the top tracker of high-frequency trading activity. The motive of the algorithm is still unclear.

The program placed orders in 25-millisecond bursts involving about 500 stocks, according to Nanex, a market data firm. The algorithm never executed a single trade, and it abruptly ended at about 10:30 a.m. ET Friday.

Why study algorithms?

Their impact is broad and far-reaching.

Internet. Web search, packet routing, distributed file sharing, ...

Biology. Human genome project, protein folding, …

Computers. Circuit layout, file system, compilers, …

Computer graphics. Movies, video games, virtual reality, …

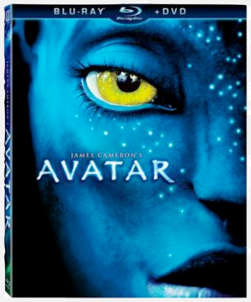
Security. Cell phones, e-commerce, voting machines, …

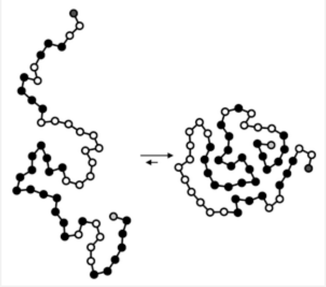
Multimedia. MP3, JPG, HDTV, song recognition, face recognition, …

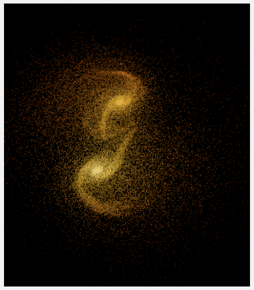
Social networks. Recommendations, dating, advertisements, …

Physics. N-body simulation, particle collision simulation, …

⋮





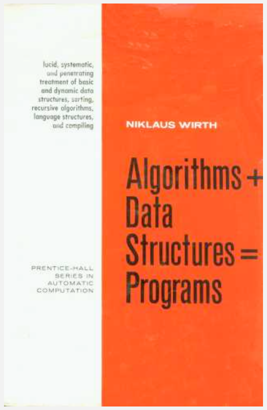


Why study algorithms?

XXVII

To become a proficient programmer.

> *"The difference between a bad programmer and a good one is whether [the programmer] considers code or data structures more important. Bad programmers worry about the code. Good programmers worry about data structures and their relationships."*
>
> *— Linus Torvalds (creator of Linux)*


lucid, systematic, and penetrating treatment of basic and dynamic data structures, sorting, recursive algorithms, language structures, and compiling
NIKLAUS WIRTH
Algorithms + Data Structures = Programs
PRENTICE-HALL SERIES IN AUTOMATIC COMPUTATION

> *" Algorithms + Data Structures = Programs. "* *— Niklaus Wirth*

Why study algorithms?

For intellectual stimulation.

Frank Nelson Cole

"On the Factorization of Large Numbers"

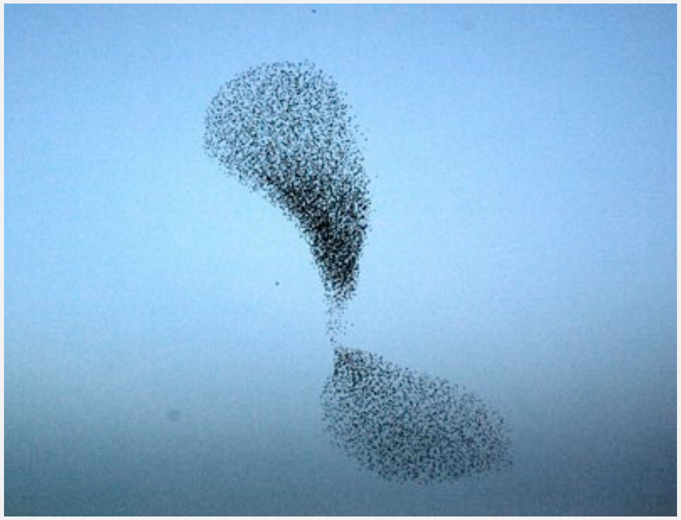
American Mathematical Society, 1903

$2^{67}-1 = 193{,}707{,}721 \times 761{,}838{,}257{,}287$

Why study algorithms?

They may unlock the secrets of life and of the universe.

Scientists are replacing mathematical models with computational models.



> *" Algorithms: a common language for nature, human, and computer. "* — *Avi Wigderson*

Why study algorithms?

For fun and profit.

Apple Computer

facebook

Cisco Systems

Google™

Nintendo®

Jane Street

IBM

Morgan Stanley

NETFLIX

Adobe™

RSA SECURITY™

DE Shaw & Co

ORACLE®

PANDORA®

Akamai

YAHOO!®

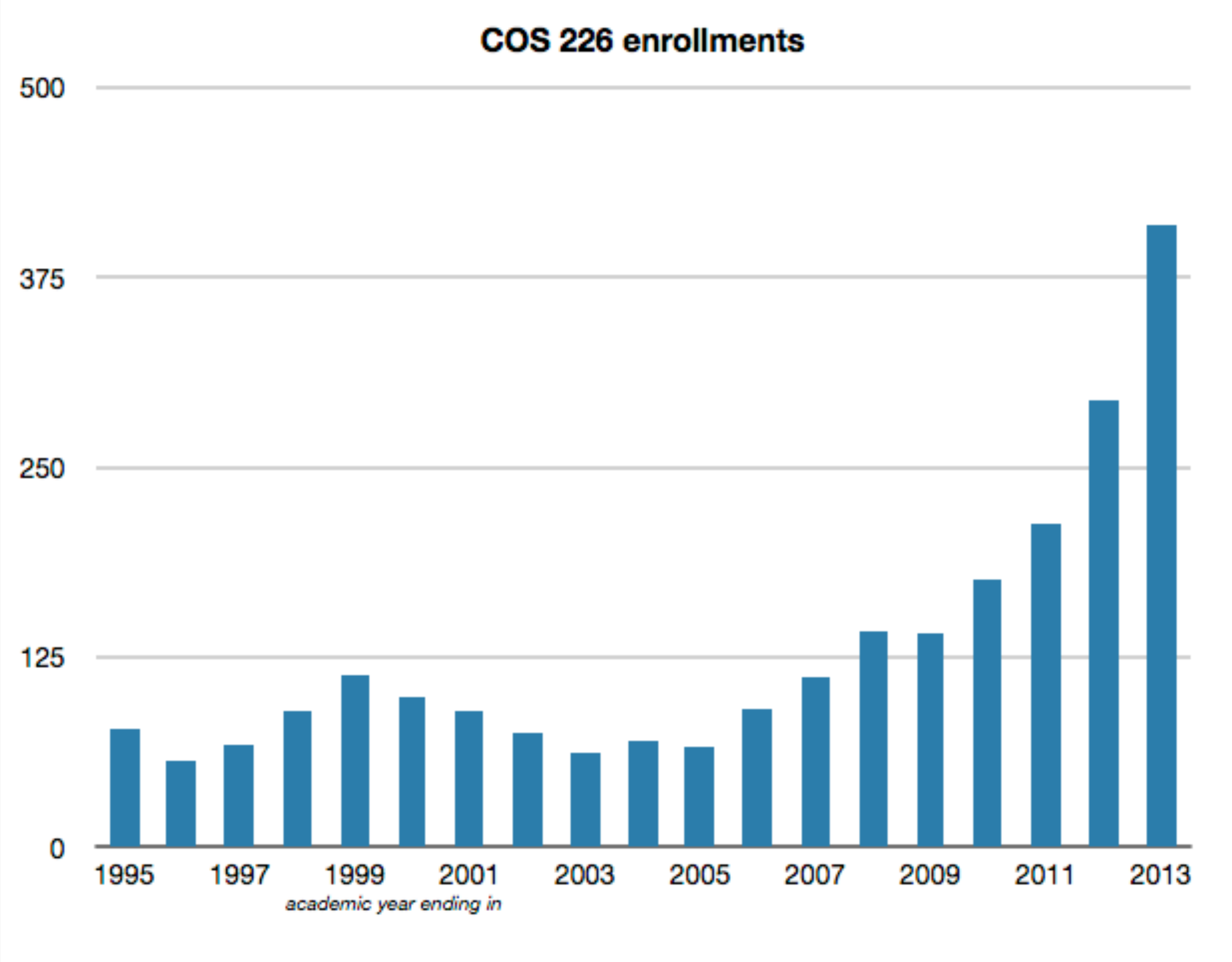
amazon.com

Microsoft®

PIXAR ANIMATION STUDIOS

Why study algorithms?

Everyone else is doing it, so why shouldn't we?

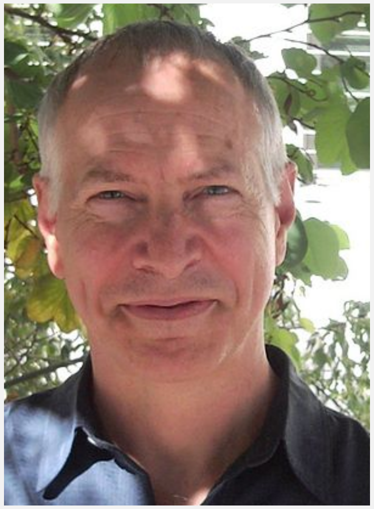


Who are you guys?

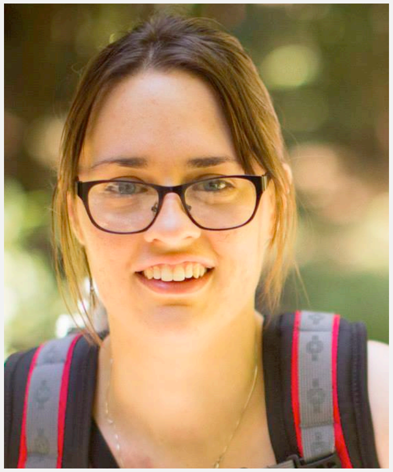
Who are we guys?

Josh Hug

Ananda **Guna**wardena



Bob Tarjan



Ruth Dannenfelser

Tengyu Ma

Deborah Varnell

Katie Edwards

The usual suspects

Lectures. Introduce new material.

Precepts. Discussion, problem-solving, background for assignments.

What	When	Where	Who	Office Hours
L01	TTh 11–12:20	Friend 101	Josh Hug	see web
P01	F 9-9:50	Friend 108	Guna †	see web
P02	F 10–10:50	Friend 108	Guna †	see web
P02A	F 10–10:50	Friend 109	Tengyu Ma	see web
P03	F 11–11:50	Friend 108	Bob Tarjan	see web
P03A	F 11–11:50	Friend 109	Deborah Varnell	see web
P04	F 12:30–1:20	Friend 108	Deborah Varnell	see web
P04A	F 12:30–1:20	Friend 109	Ruth Dannenfelser	see web

† lead preceptor

Where to get help?

Piazza. Online discussion forum.

- Low latency, low bandwidth.
- Mark solution-revealing questions as private.
- Course announcements.

piazza

http://www.piazza.com/class#fall2013/cos226

Office hours.

- High bandwidth, high latency.
- See web for schedule.

http://www.princeton.edu/~cos226

Computing laboratory.

- Undergrad lab TAs in Friend 017.
- For help with debugging.
- See web for schedule.

http://www.princeton.edu/~cos226

Coursework and grading

Programming assignments. 45%

- Due on Wednesdays at 11:00 pm via electronic submission.
- 4 free late days. Lose 10% for each late day thereafter.
- See web for full collaboration and lateness policy.

Exercises. 10%

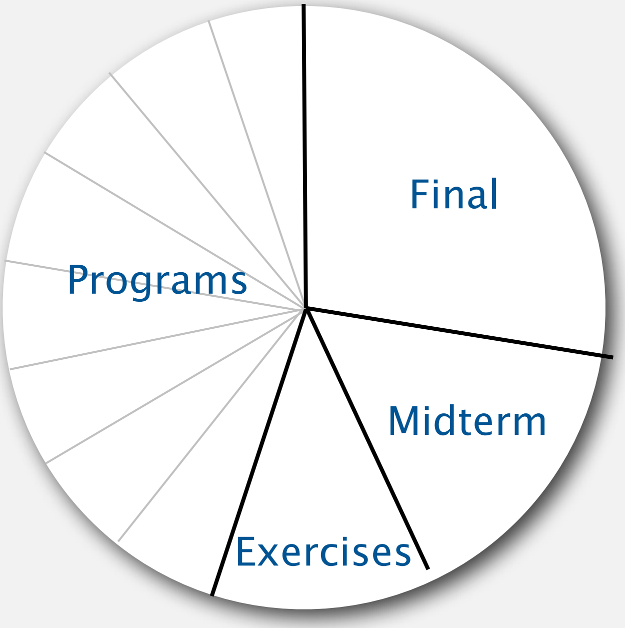
- Due on Sundays at 11pm in Blackboard.

Exams. 15% + 30%

- Midterm (in class on Tuesday, October 22).
- Final (to be scheduled by Registrar).

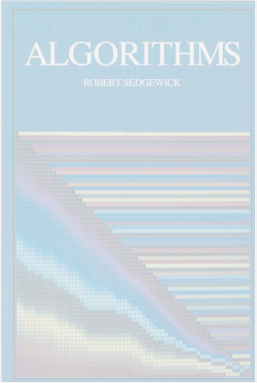
Staff discretion. To adjust borderline cases.

- Report errata.
- Contribute to Piazza discussions.
- Attend and participate in precept/lecture.
- Answering in lecture-questions using a device.

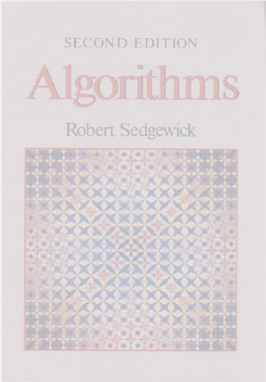


Resources (textbook)

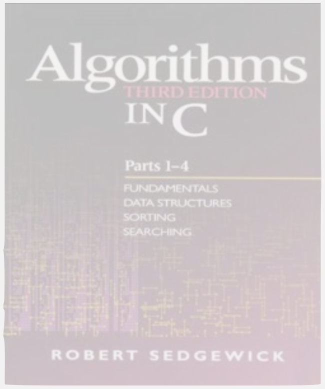
Required reading. Algorithms 4th edition by R. Sedgewick and K. Wayne, Addison-Wesley Professional, 2011, ISBN 0-321-57351-X.



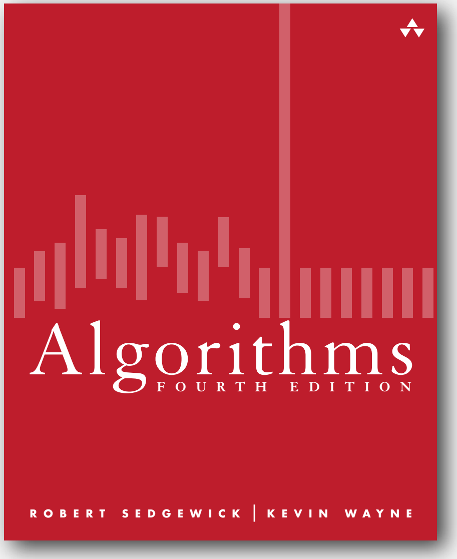
1st edition (1982)



2nd edition (1988)



3rd edition (1997)



Available in hardcover and Kindle.

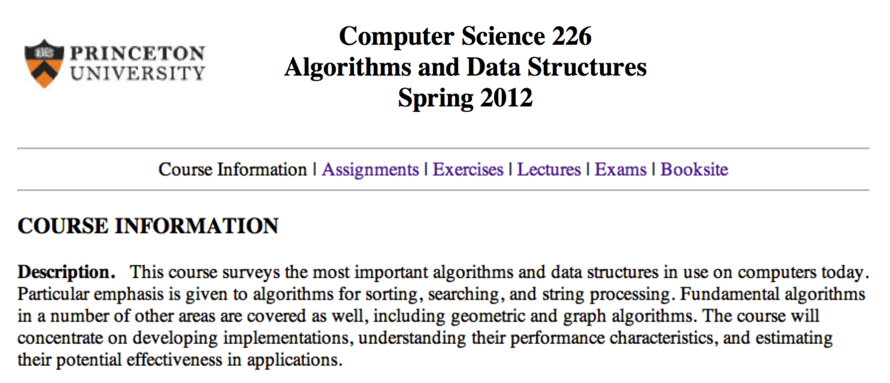
- Online: Amazon ($60 to buy), Chegg ($40 to rent), ...
- Brick-and-mortar: Labyrinth Books (122 Nassau St). ←
- On reserve: Engineering library.

30% discount with PU student ID

Resources (web)

Course content.

- Course info.
- Programming assignments.
- Exercises.
- Lecture slides.
- Exam archive.
- Submit assignments.



PRINCETON UNIVERSITY

Computer Science 226
Algorithms and Data Structures
Spring 2012

Course Information | Assignments | Exercises | Lectures | Exams | Booksite

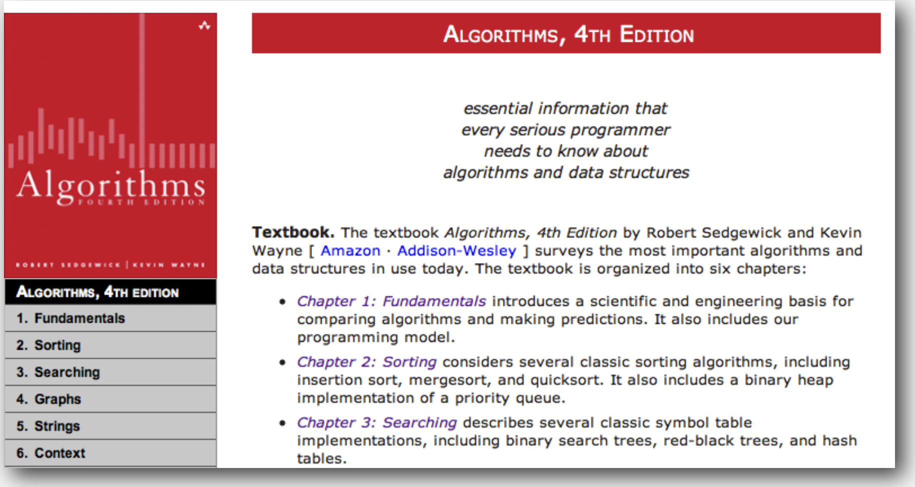
COURSE INFORMATION

Description. This course surveys the most important algorithms and data structures in use on computers today. Particular emphasis is given to algorithms for sorting, searching, and string processing. Fundamental algorithms in a number of other areas are covered as well, including geometric and graph algorithms. The course will concentrate on developing implementations, understanding their performance characteristics, and estimating their potential effectiveness in applications.

http://www.princeton.edu/~cos226

Booksites.

- Brief summary of content.
- Download code from book.



ALGORITHMS, 4TH EDITION

*essential information that
every serious programmer
needs to know about
algorithms and data structures*

Textbook. The textbook *Algorithms, 4th Edition* by Robert Sedgewick and Kevin Wayne [Amazon · Addison-Wesley] surveys the most important algorithms and data structures in use today. The textbook is organized into six chapters:

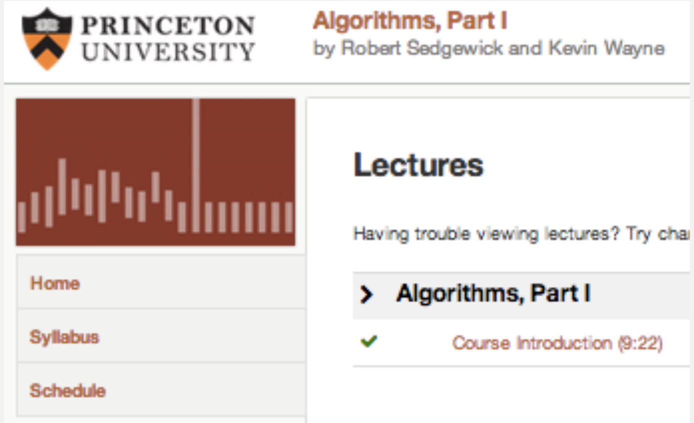
- *Chapter 1: Fundamentals* introduces a scientific and engineering basis for comparing algorithms and making predictions. It also includes our programming model.
- *Chapter 2: Sorting* considers several classic sorting algorithms, including insertion sort, mergesort, and quicksort. It also includes a binary heap implementation of a priority queue.
- *Chapter 3: Searching* describes several classic symbol table implementations, including binary search trees, red-black trees, and hash tables.

http://www.algs4.princeton.edu

Resources (Coursera) and Flipped Lectures

Coursera Course

- Videos by Bob Sedgewick.
  - Nearly same content as ours.
- Don't submit assignments!
  - Violates course policy.



PRINCETON UNIVERSITY

Algorithms, Part I

by Robert Sedgewick and Kevin Wayne

Lectures

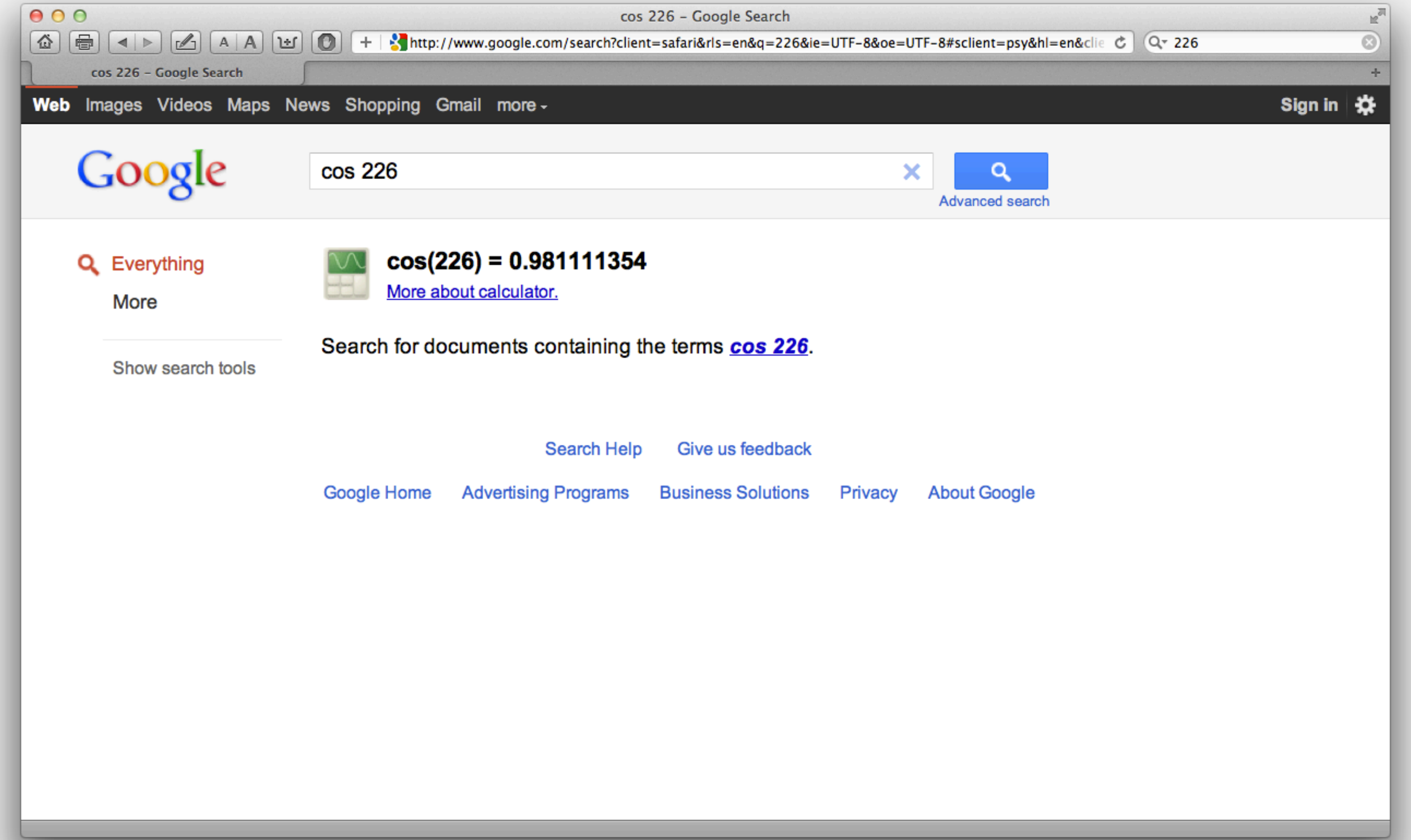
Having trouble viewing lectures? Try cha

https://class.coursera.org/algs4partI-003/class

Flipped Lectures

- Special large-room format office hours (time to be scheduled)
  - Me / Guna solving hard problems
  - Old exam problems
  - Open Q&A
- Alternative or supplement to in-class lectures.
- Not required. Attendance not tracked.

Resources (web)



cos 226

cos(226) = 0.981111354

More about calculator.

Search for documents containing the terms *cos 226*.

http://www.princeton.edu/~cos226

Resources (web)

About 236,000,000 results (0.18 seconds)

[Area codes 519 and **226** - Wikipedia, the free encyclopedia]
The **226** area code was first proposed as a result of an NPA exhaustion study conducted in the 1990s. The issue was raised with the CRTC by telecommunications ...
en.wikipedia.org/wiki/Area_codes_519_and_**226** - Cached - Similar

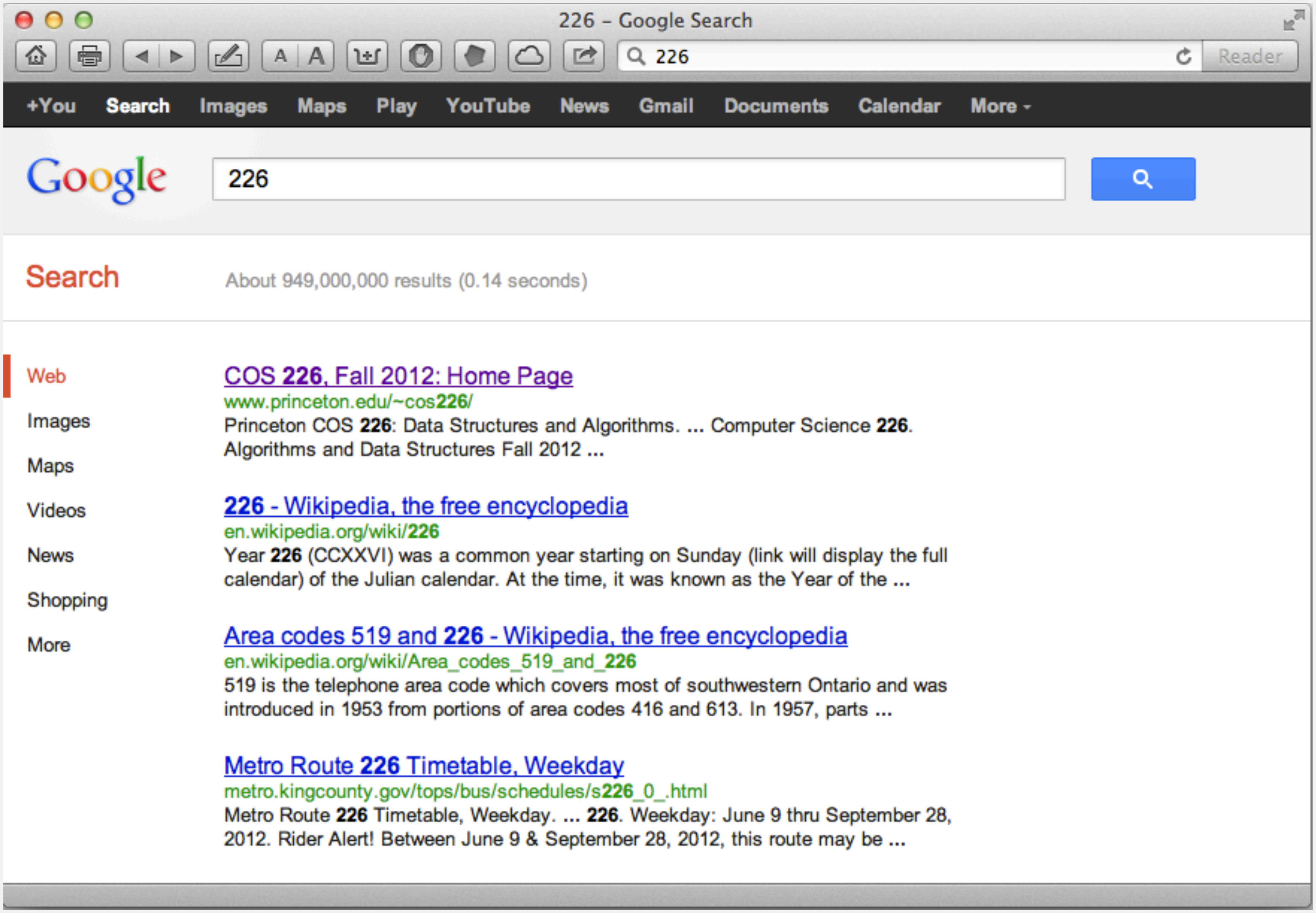
[**226** - Wikipedia, the free encyclopedia]
226. From Wikipedia, the free encyclopedia. Jump to: navigation, search. This article is about the year **226**. For the number **226**, see **226** (number). ...
en.wikipedia.org/wiki/**226** - Cached - Similar

[COS **226**, Fall 2010: Home Page]
Princeton COS **226**: Data Structures and Algorithms. ... Computer Science **226**. Algorithms and Data Structures Fall 2010 ...
www.princeton.edu/~cos**226**/ - Cached - Similar

Images for **226** - Report images

http://www.princeton.edu/~cos226

Resources (web)



226

Search

About 949,000,000 results (0.14 seconds)

COS **226**, Fall 2012: Home Page
www.princeton.edu/~cos**226**/
Princeton COS **226**: Data Structures and Algorithms. ... Computer Science **226**. Algorithms and Data Structures Fall 2012 ...

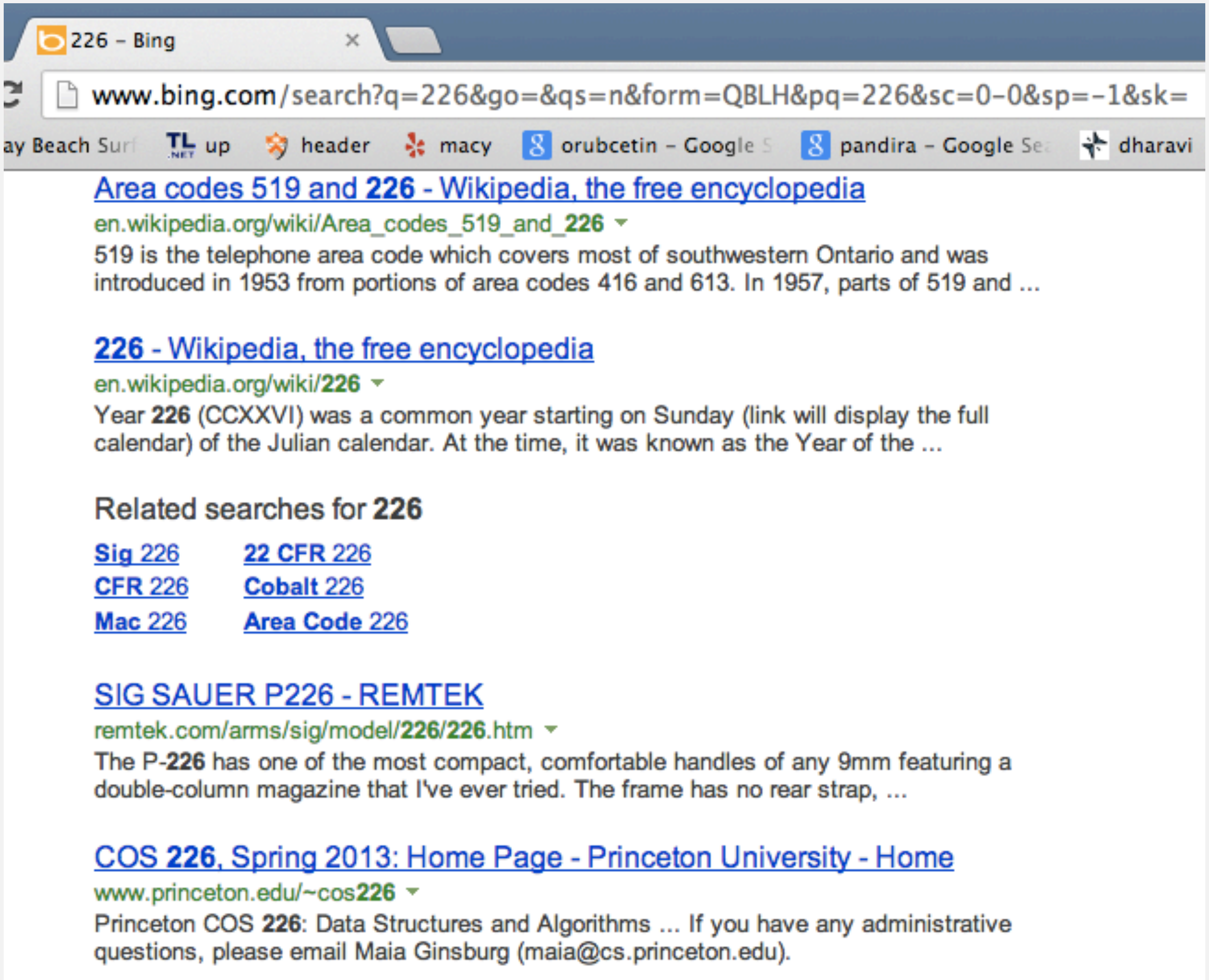
226 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**226**
Year **226** (CCXXVI) was a common year starting on Sunday (link will display the full calendar) of the Julian calendar. At the time, it was known as the Year of the ...

Area codes 519 and **226** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Area_codes_519_and_**226**
519 is the telephone area code which covers most of southwestern Ontario and was introduced in 1953 from portions of area codes 416 and 613. In 1957, parts ...

Metro Route **226** Timetable, Weekday
metro.kingcounty.gov/tops/bus/schedules/s**226**_0_.html
Metro Route **226** Timetable, Weekday. ... **226**. Weekday: June 9 thru September 28, 2012. Rider Alert! Between June 9 & September 28, 2012, this route may be ...

http://www.princeton.edu/~cos226

Resources (web)



Area codes 519 and **226** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Area_codes_519_and_**226**
519 is the telephone area code which covers most of southwestern Ontario and was introduced in 1953 from portions of area codes 416 and 613. In 1957, parts of 519 and ...

226 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**226**
Year **226** (CCXXVI) was a common year starting on Sunday (link will display the full calendar) of the Julian calendar. At the time, it was known as the Year of the ...

Related searches for **226**

Sig 226	**22 CFR** 226
CFR 226	**Cobalt** 226
Mac 226	**Area Code** 226

SIG SAUER P226 - REMTEK
remtek.com/arms/sig/model/**226**/**226**.htm
The P-**226** has one of the most compact, comfortable handles of any 9mm featuring a double-column magazine that I've ever tried. The frame has no rear strap, ...

COS **226**, Spring 2013: Home Page - Princeton University - Home
www.princeton.edu/~cos**226**
Princeton COS **226**: Data Structures and Algorithms ... If you have any administrative questions, please email Maia Ginsburg (maia@cs.princeton.edu).

A note on cheating

Cheating

- Don't.
- More than two dozen cases last semester in lower division CS courses.
- We possess and utilize highly advanced tools to detect plagiarism.
- Most likely penalty is a one year-suspension.
  - Copying code.
  - Looking at other student's (past or present) code.
  - Giving your code to someone else (present or future).
  - Submitting to the Coursera autograder.
  - COS226 staff have no discretion!

What's ahead?

Lecture 1. [today] Union find.

Lecture 2. [Next Tuesday] Analysis of algorithms.

Precept 1. [Friday] Meets this week.

Exercise 1. Due via Bb submission at 11pm on Sunday, September 15th.

Assignment 1. Due via electronic submission at 11:59pm on Wednesday, September 18th. Pro tip: Start early (after precept tomorrow).

Right course? See me.

Placed out of COS 126? Review Sections 1.1–1.2 of Algorithms, 4th edition (includes command-line interface and our I/O libraries).

Not registered? Go to any precept this week [only if not registered!].

Change precept? Use SCORE. ← Will set up system to swap next week

Algorithms
FOURTH EDITION
ROBERT SEDGEWICK | KEVIN WAYNE

http://algs4.cs.princeton.edu

1.5 UNION-FIND

- *dynamic connectivity*
- *quick find*
- *quick union*
- *improvements*
- *applications*

Subtext of today's lecture (and this course)

Steps to developing a usable algorithm.

- Model the problem.
- Find an algorithm to solve it.
- Fast enough? Fits in memory?
- If not, figure out why.
- Find a way to address the problem.
- Iterate until satisfied.

The scientific method.

Mathematical analysis.

Algorithms

Robert Sedgewick | Kevin Wayne

http://algs4.cs.princeton.edu

1.5 Union-Find

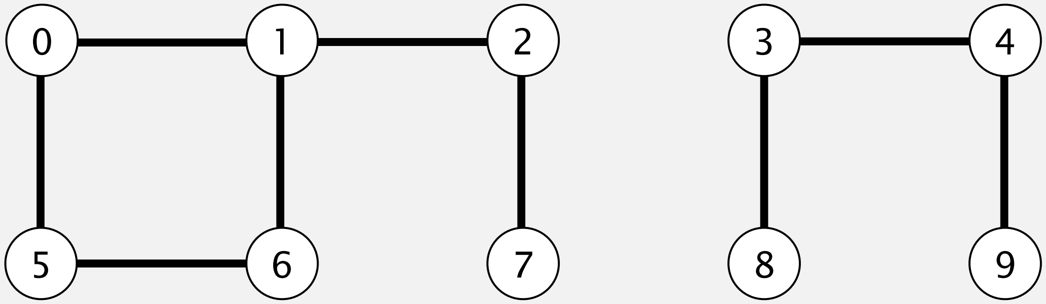
- *dynamic connectivity*
- *quick find*
- *quick union*
- *improvements*
- *applications*

Dynamic connectivity

Given a set of N objects.

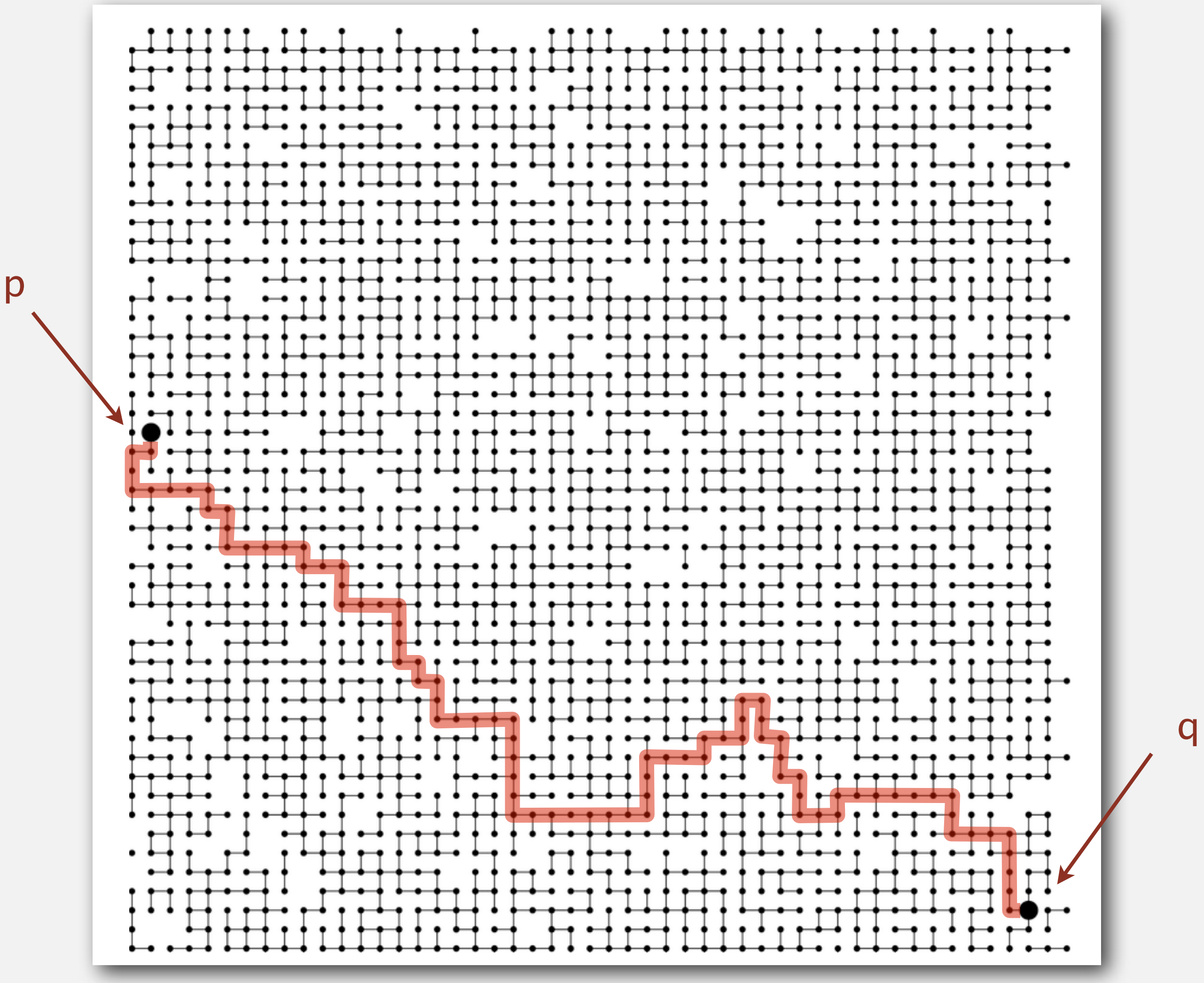
- Union command: connect two objects.
- Find/connected query: is there a path connecting the two objects?

```
union(4, 3)
union(3, 8)
union(6, 5)
union(9, 4)
union(2, 1)
connected(0, 7)    ✗
connected(8, 9)    ✔
union(5, 0)
union(7, 2)
union(6, 1)
union(1, 0)
connected(0, 7)    ✔
```



Connectivity example

Q. Is there a path connecting p and q ?



A. Yes.

Modeling the objects

Applications involve manipulating objects of all types.

- Pixels in a digital photo.
- Computers in a network.
- Friends in a social network.
- Transistors in a computer chip.
- Elements in a mathematical set.
- Variable names in Fortran program.
- Metallic sites in a composite system.

When programming, convenient to name objects 0 to N –1.

- Use integers as array index.
- Suppress details not relevant to union-find.

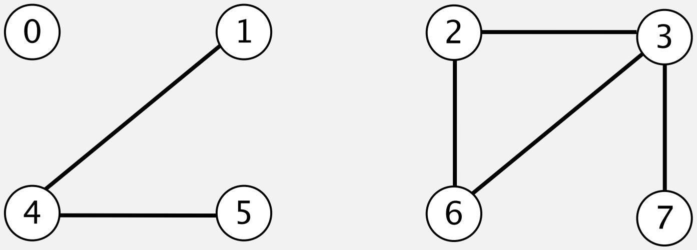
can use symbol table to translate from site names to integers: stay tuned (Chapter 3)

Modeling the connections

We assume "is connected to" is an equivalence relation:

- Reflexive: p is connected to p.
- Symmetric: if p is connected to q, then q is connected to p.
- Transitive: if p is connected to q and q is connected to r, then p is connected to r.

Connected components. Maximal set of objects that are mutually connected.



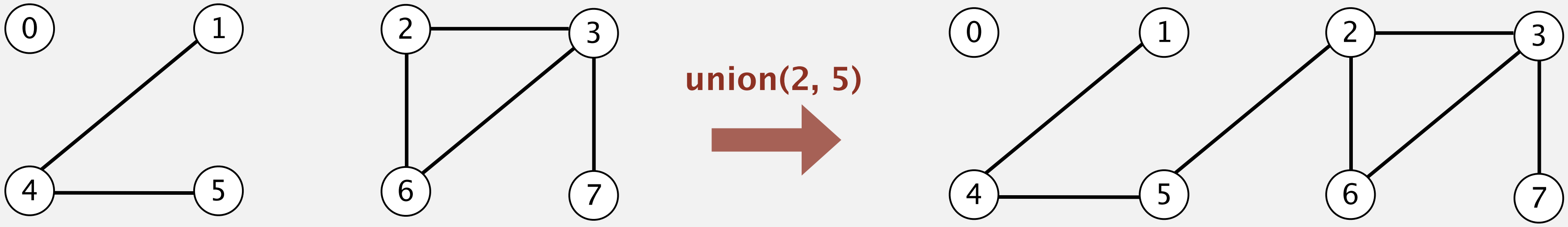
{ 0 } { 1 4 5 } { 2 3 6 7 }

3 connected components

Implementing the operations

Find query. Check if two objects are in the same component.

Union command. Replace components containing two objects with their union.



{ 0 } { 1 4 5 } { 2 3 6 7 }

3 connected components

{ 0 } { 1 2 3 4 5 6 7 }

2 connected components

Union-find data type (API)

Goal. Design efficient data structure for union-find.

- Number of objects N can be huge.
- Number of operations M can be huge.
- Find queries and union commands may be intermixed.

public class UF		
	`UF(int N)`	*initialize union-find data structure with N objects (0 to $N - 1$)*
`void`	`union(int p, int q)`	*add connection between p and q*
`boolean`	`connected(int p, int q)`	*are p and q in the same component?*
`int`	`find(int p)`	*component identifier for p (0 to $N - 1$)*
`int`	`count()`	*number of components*

Dynamic-connectivity client

- Read in number of objects N from standard input.
- Repeat:
  - read in pair of integers from standard input
  - if they are not yet connected, connect them and print out pair

```
public static void main(String[] args)
{
   int N = StdIn.readInt();
   UF uf = new UF(N);
   while (!StdIn.isEmpty())
   {
      int p = StdIn.readInt();
      int q = StdIn.readInt();
      if (!uf.connected(p, q))
      {
         uf.union(p, q);
         StdOut.println(p + " " + q);
      }
   }
}
```

```
% more tinyUF.txt
10
4 3
3 8
6 5
9 4
2 1
8 9
5 0
7 2
6 1
1 0
6 7
```

Algorithms

Robert Sedgewick | Kevin Wayne

http://algs4.cs.princeton.edu

1.5 Union-Find

- *dynamic connectivity*
- *quick find*
- *quick union*
- *improvements*
- *applications*

Quick-find [eager approach]

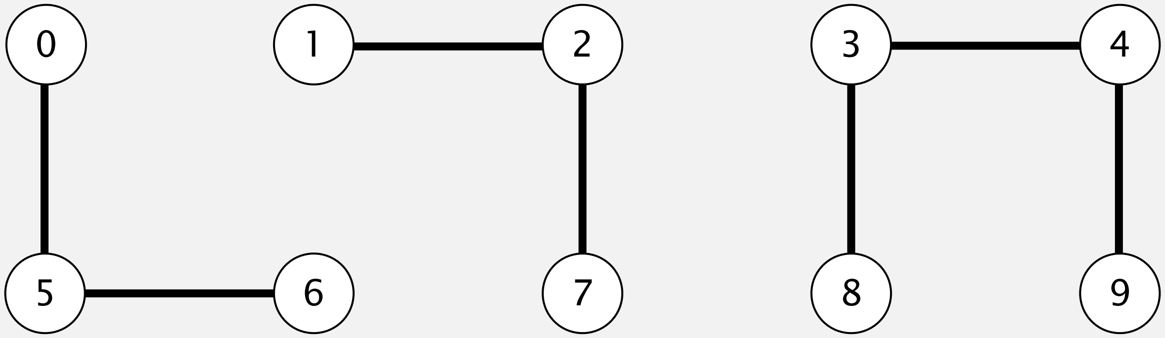
Data structure.

- Integer array `id[]` of size `N`.
- Interpretation: `p` and `q` are connected iff they have the same id

if and only if

	0	1	2	3	4	5	6	7	8	9
`id[]`	0	1	1	8	8	0	0	1	8	8

0, 5 and 6 are connected
1, 2, and 7 are connected
3, 4, 8, and 9 are connected



Quick-find [eager approach]

Data structure.

- Integer array `id[]` of size `N`.
- Interpretation: `p` and `q` are connected iff they have the same id.

	0	1	2	3	4	5	6	7	8	9
`id[]`	0	1	1	8	8	0	0	1	8	8

Find. id of `p` gives its component.

If `p` and `q` have the same id, they are connected.

`id[6] = 0`; `id[1] = 1`

6 and 1 are not connected

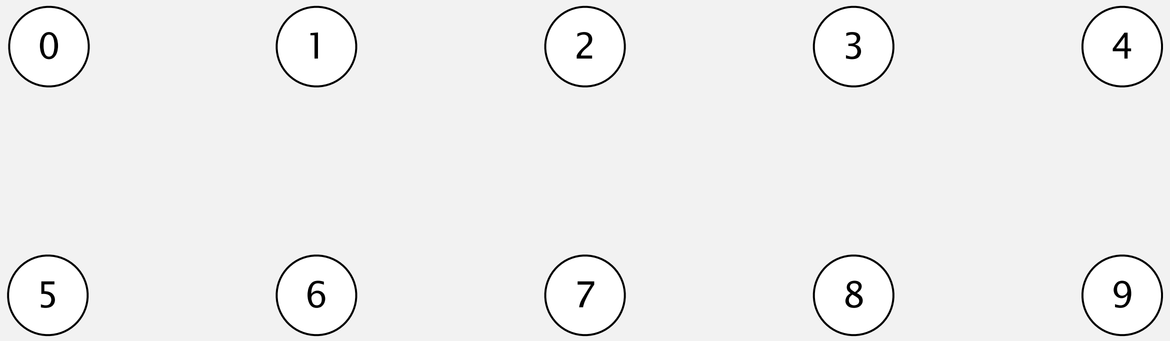
Union. To merge components containing `p` and `q`, change all entries whose id equals `id[p]` to `id[q]`.

	0	1	2	3	4	5	6	7	8	9
`id[]`	1	1	1	8	8	1	1	1	8	8

after union of 6 and 1

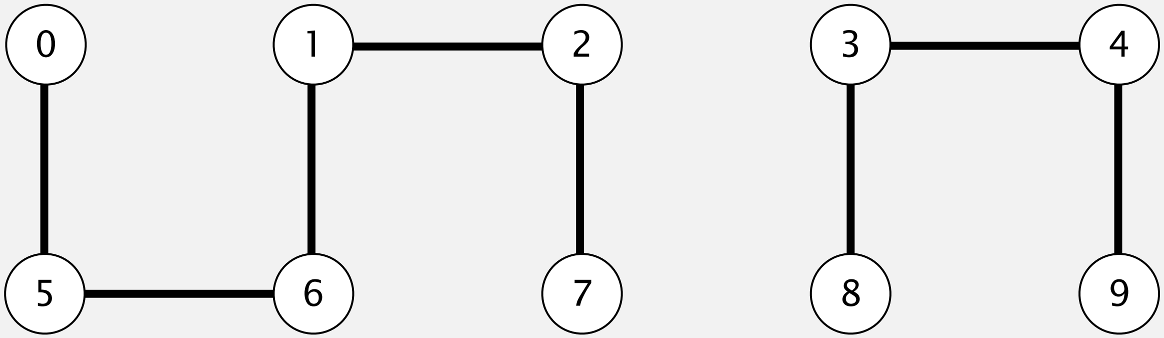
problem: many values can change

Quick-find demo



	0	1	2	3	4	5	6	7	8	9
id[]	0	1	2	3	4	5	6	7	8	9

Quick-find demo



	0	1	2	3	4	5	6	7	8	9
id[]	1	1	1	8	8	1	1	1	8	8

Quick-find: Java implementation

```
public class QuickFindUF
{
   private int[] id;

   public QuickFindUF(int N)
   {

   }

   public boolean connected(int p, int q)
   { }

   public void union(int p, int q)
   {
   }
}
```

Quick-find: Java implementation

```
public class QuickFindUF
{
   private int[] id;

   public QuickFindUF(int N)
   {
      id = new int[N];
      for (int i = 0; i < N; i++)
         id[i] = i;
   }

   public boolean connected(int p, int q)
   {  return id[p] == id[q];  }

   public void union(int p, int q)
   {
      int pid = id[p];
      int qid = id[q];
      for (int i = 0; i < id.length; i++)
         if (id[i] == pid) id[i] = qid;
   }
}
```

set id of each object to itself
(N array accesses)

check whether p and q
are in the same component
(2 array accesses)

change all entries with id[p] to id[q]
(at most 2N + 2 array accesses)

Quick-find is too slow

Cost model. Number of array accesses (for read or write).

algorithm	initialize	union	find
quick-find	N	N	1

order of growth of number of array accesses

quadratic

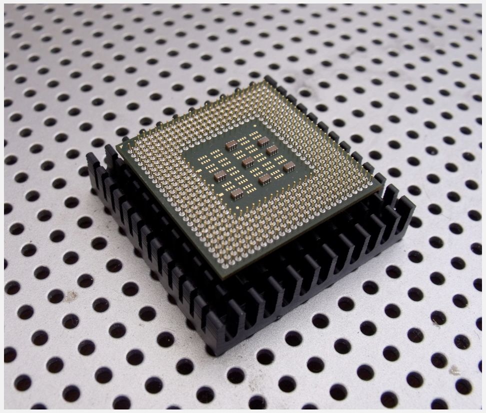
Union is too expensive. It takes N^2 array accesses to process a sequence of N union commands on N objects.

Quadratic algorithms do not scale

Rough standard (for now).

- 10^9 operations per second.
- 10^9 words of main memory.
- Touch all words in approximately 1 second.

a truism (roughly) since 1950!

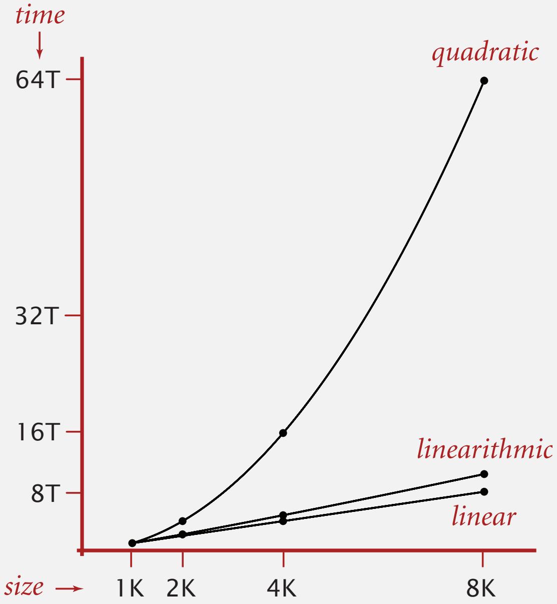


Ex. Huge problem for quick-find.

- 10^9 union commands on 10^9 objects.
- Quick-find takes more than 10^{18} operations.
- 30+ years of computer time!

Quadratic algorithms don't scale with technology.

- New computer may be 10x as fast.
- But, has 10x as much memory ⇒ want to solve a problem that is 10x as big.
- With quadratic algorithm, takes 10x as long!



http://algs4.cs.princeton.edu

1.5 Union-Find

- dynamic connectivity
- quick find
- quick union
- improvements
- applications

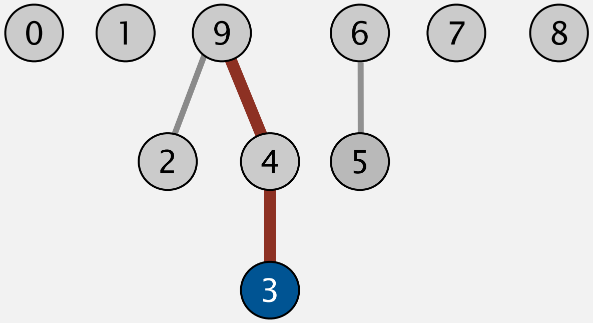
Quick-union [lazy approach]

Data structure.

- Integer array `id[]` of size `N`.
- Interpretation: `id[i]` is parent of `i`.
- Root of `i` is `id[id[id[...id[i]...]]]`.

keep going until it doesn't change
(algorithm ensures no cycles)

	0	1	2	3	4	5	6	7	8	9
id[]	0	1	9	4	9	6	6	7	8	9



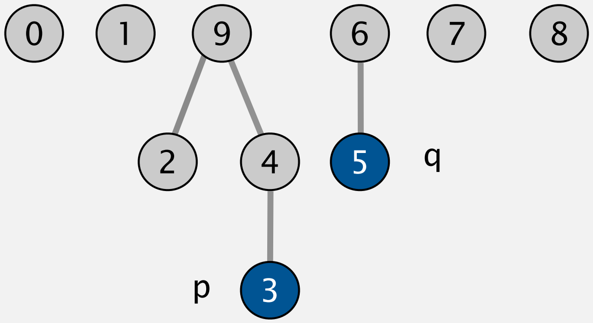
root of 3 is 9

Quick-union [lazy approach]

Data structure.

- Integer array `id[]` of size `N`.
- Interpretation: `id[i]` is parent of `i`.
- Root of `i` is `id[id[id[...id[i]...]]]`.

	0	1	2	3	4	5	6	7	8	9
id[]	0	1	9	4	9	6	6	7	8	9



root of 3 is 9

root of 5 is 6

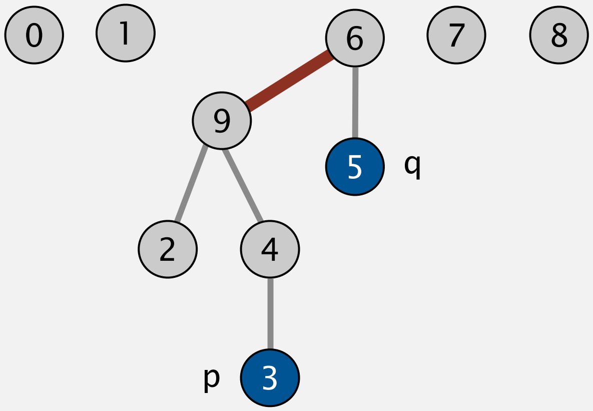
3 and 5 are not connected

Find. Check if `p` and `q` have the same root.

Union. To merge components containing `p` and `q`, set the id of `p`'s root to the id of `q`'s root.

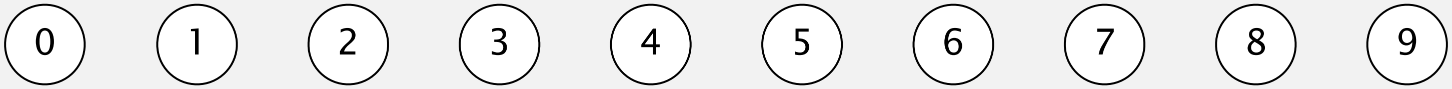
	0	1	2	3	4	5	6	7	8	9
id[]	0	1	9	4	9	6	6	7	8	6

only one value changes



Quick-union demo

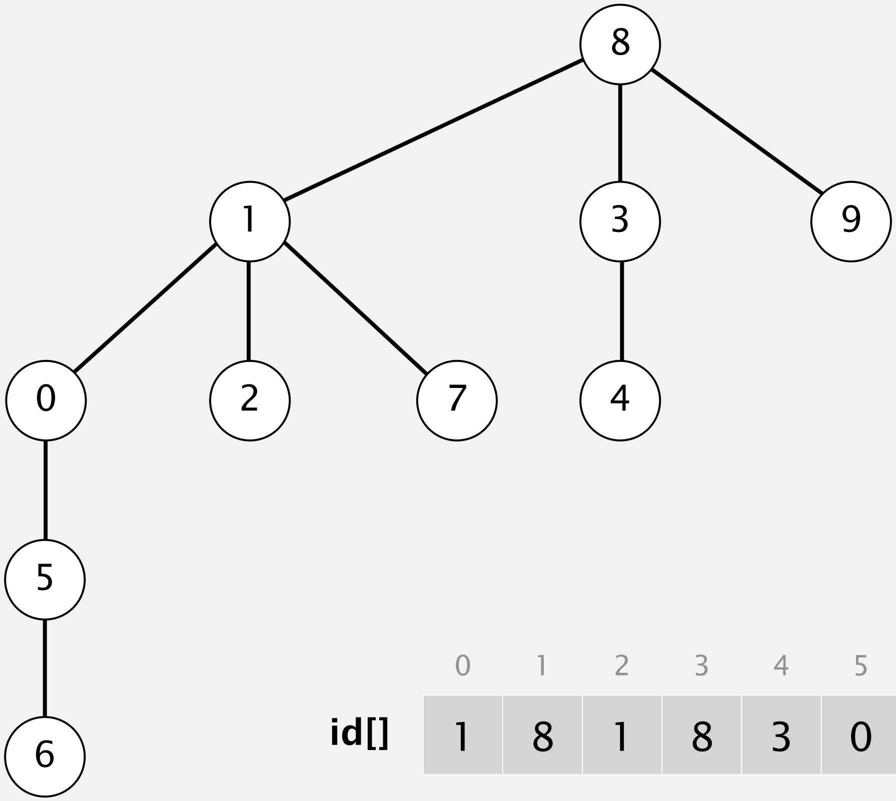
▶



	0	1	2	3	4	5	6	7	8	9
id[]	0	1	2	3	4	5	6	7	8	9

Quick-union demo

Question: Worst case tree depth? Best Case?



	0	1	2	3	4	5	6	7	8	9
id[]	1	8	1	8	3	0	5	1	8	8

Quick-union: Java implementation

```java
public class QuickUnionUF
{
   private int[] id;

   public QuickUnionUF(int N)
   {
      id = new int[N];
      for (int i = 0; i < N; i++) id[i] = i;
   }

   private int root(int i)
   {
      while (i != id[i]) i = id[i];
      return i;
   }

   public boolean connected(int p, int q)
   {
      return root(p) == root(q);
   }

   public void union(int p, int q)
   {
      int i = root(p);
      int j = root(q);
      id[i] = j;
   }
}
```

set id of each object to itself
(N array accesses)

chase parent pointers until reach root
(depth of i array accesses)

check if p and q have same root
(depth of p and q array accesses)

change root of p to point to root of q
(depth of p and q array accesses)

Quick-union is also too slow

Cost model. Number of array accesses (for read or write).

algorithm	initialize	union	find
quick-find	N	N	1
quick-union	N	N[†]	N

worst case

[†] includes cost of finding roots

Quick-find defect.

- Union too expensive (N array accesses).
- Trees are flat, but too expensive to keep them flat.

Quick-union defect.

- Trees can get tall.
- Find too expensive (could be N array accesses).

http://algs4.cs.princeton.edu

1.5 Union-Find

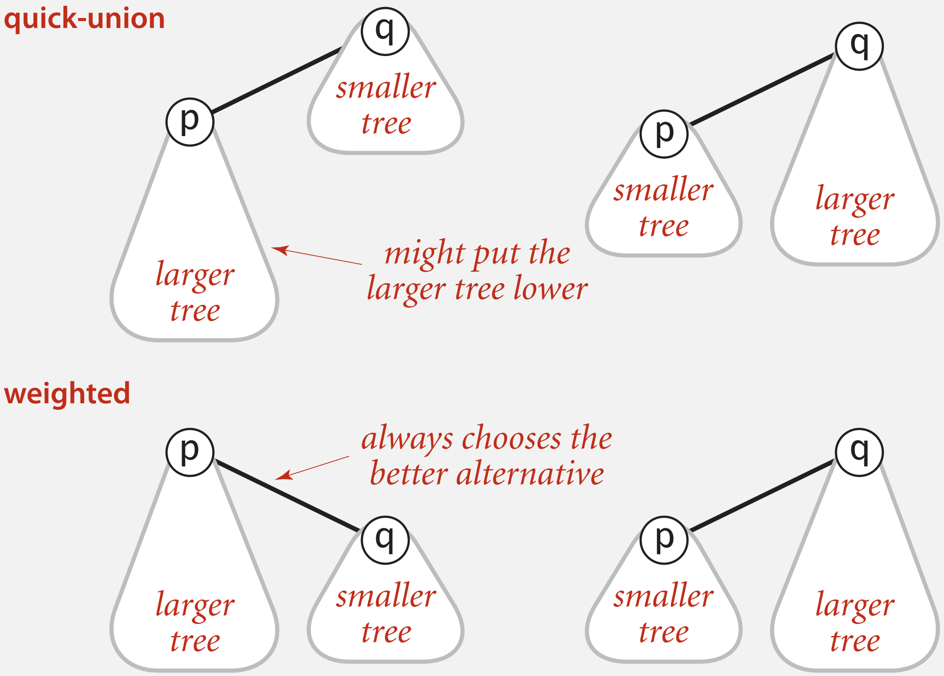
- *dynamic connectivity*
- *quick find*
- *quick union*
- *improvements*
- *applications*

Improvement 1: weighting

Weighted quick-union.

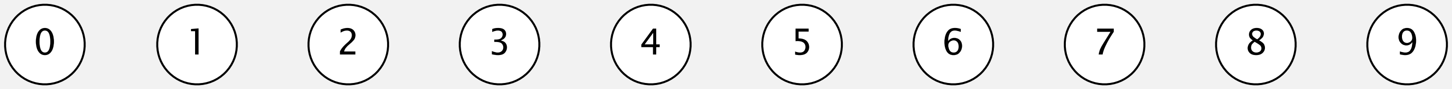
- Modify quick-union to avoid tall trees.
- Keep track of size of each tree (number of objects).
- Balance by linking root of smaller tree to root of larger tree.

reasonable alternatives:
union by height or "rank"



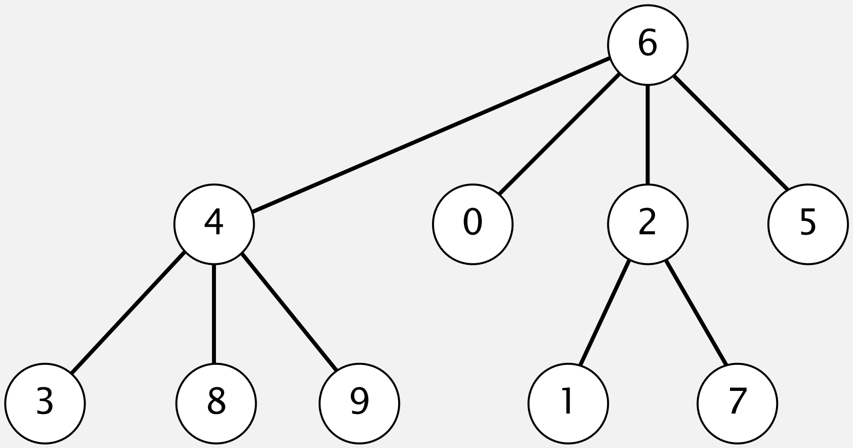
In short: Keep **union** from unnecessarily lengthening the tree.

Weighted quick-union demo



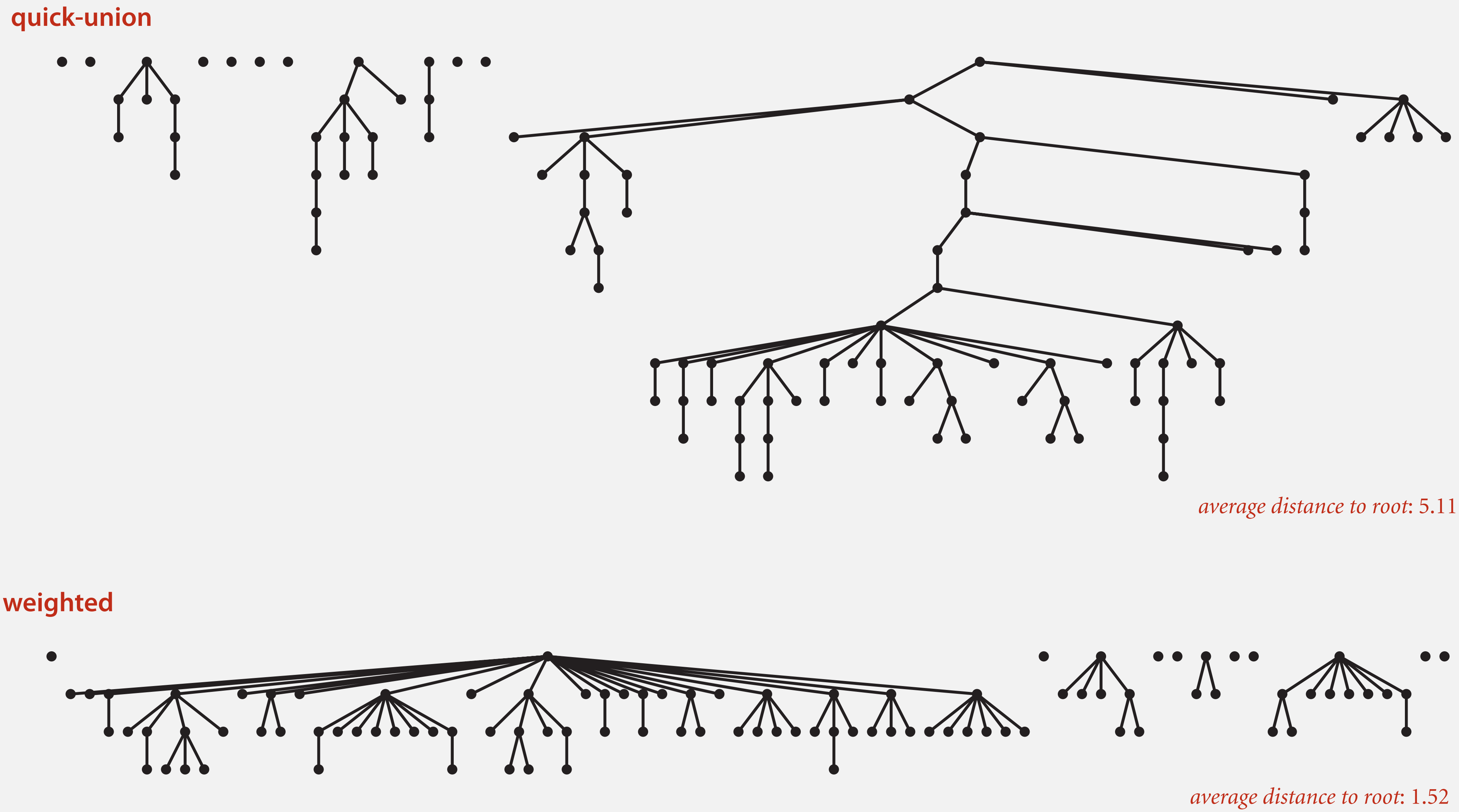
	0	1	2	3	4	5	6	7	8	9
id[]	0	1	2	3	4	5	6	7	8	9

Weighted quick-union demo



	0	1	2	3	4	5	6	7	8	9
id[]	6	2	6	4	6	6	6	2	4	4

Quick-union and weighted quick-union example



quick-union

average distance to root: 5.11

weighted

average distance to root: 1.52

Quick-union and weighted quick-union (100 sites, 88 `union()` operations)

Weighted quick-union: Java implementation

Data structure. Same as quick-union, but maintain extra array `sz[i]` to count number of objects in the tree rooted at `i`.

Find. Identical to quick-union.

```
return root(p) == root(q);
```

Union. Modify quick-union to:

- Link root of smaller tree to root of larger tree.
- Update the `sz[]` array.

```
int i = root(p);
int j = root(q);
if  (sz[i] < sz[j]) { id[i] = j; sz[j] += sz[i]; }
else                { id[j] = i; sz[i] += sz[j]; }
```

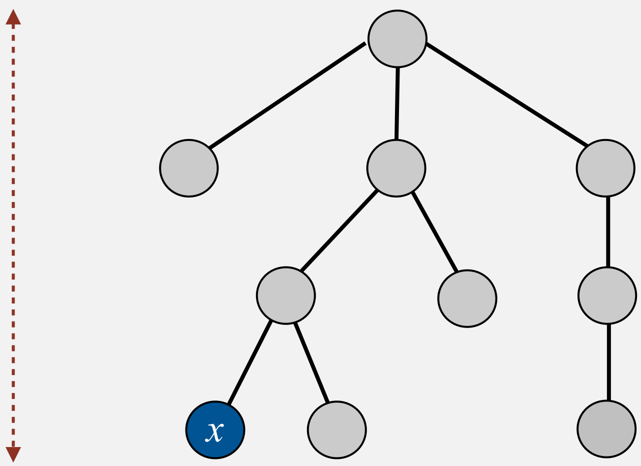
Weighted quick-union analysis

Running time.

- Find: takes time proportional to depth of p and q.
- Union: takes constant time, given roots.

lg = base-2 logarithm

Proposition. Depth of any node x is at most $\lg N$.



N = 10
depth(x) = 3 ≤ lg N

Weighted quick-union analysis

Running time.

- Find: takes time proportional to depth of p and q.
- Union: takes constant time, given roots.

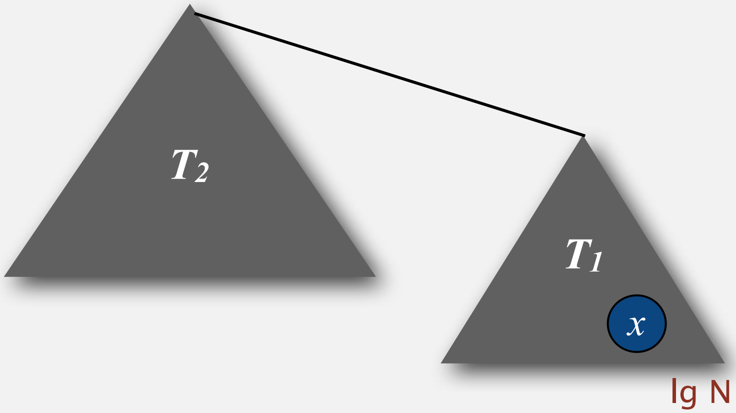
lg = base-2 logarithm

Proposition. Depth of any node x is at most $\lg N$.

Pf. When does depth of x increase?

Increases by 1 when tree T_1 containing x is merged into another tree T_2.

- The size of the tree containing x at least doubles since $|T_2| \geq |T_1|$.
- Size of tree containing x can double at most $\lg N$ times. Why?



lg N

Size of tree containing x = 2*2*2*... = N

Weighted quick-union analysis

Running time.

- Find: takes time proportional to depth of p and q.
- Union: takes constant time, given roots.

Proposition. Depth of any node x is at most $\lg N$.

algorithm	initialize	union	connected
quick-find	N	N	1
quick-union	N	N †	N
weighted QU	N	lg N †	lg N

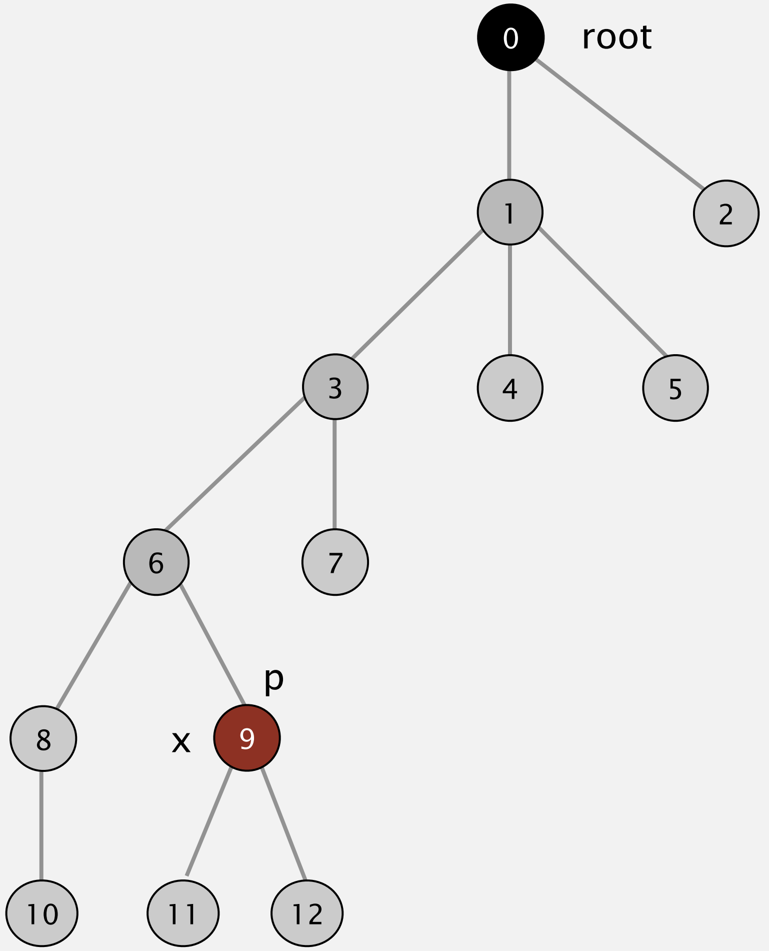
† includes cost of finding roots

Q. Stop at guaranteed acceptable performance?

A. No, easy to improve further.

Improvement 2: path compression

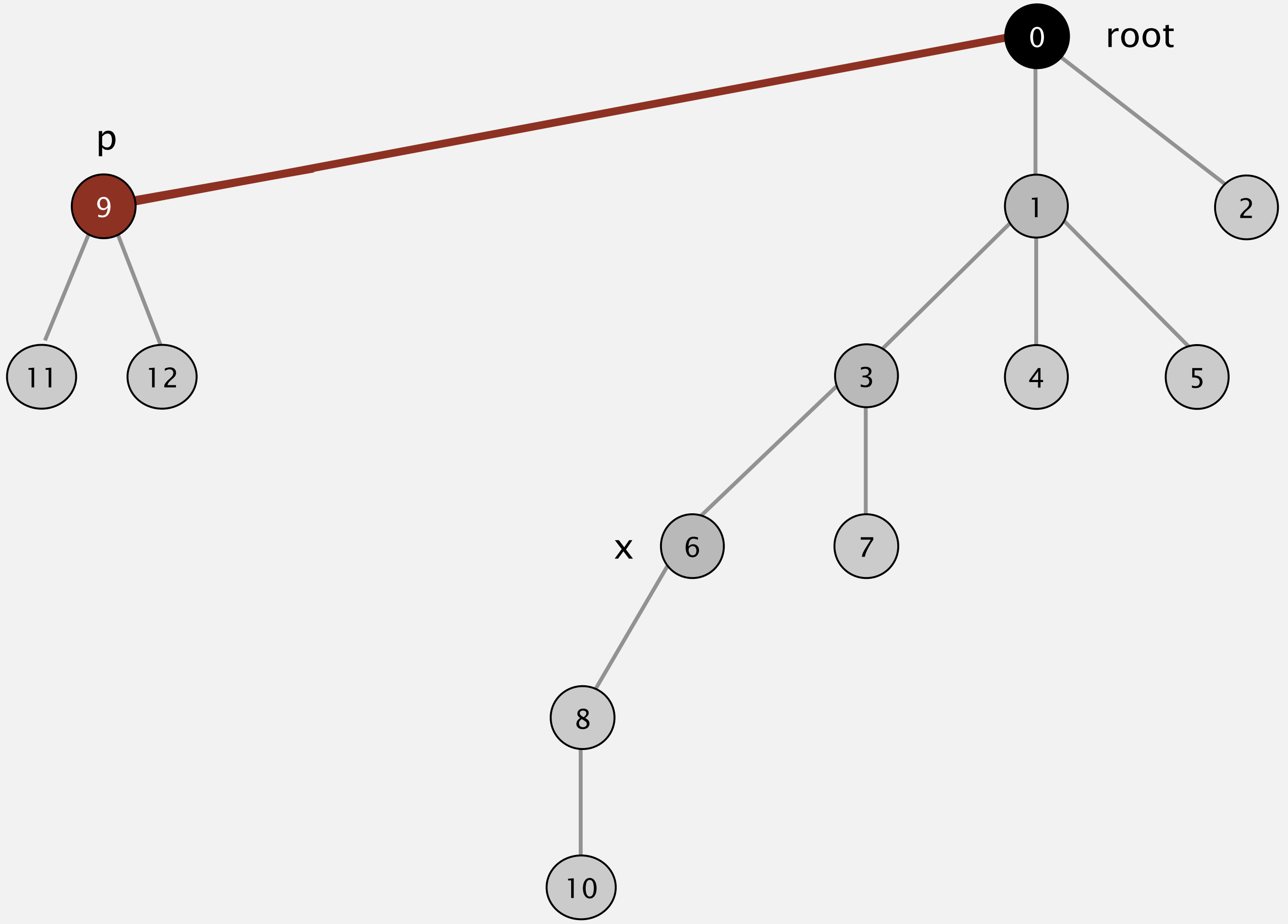
Quick union with path compression. Just after computing the root of p, set the id of each examined node to point to that root.



In short: Give **find** a side job compressing the tree.

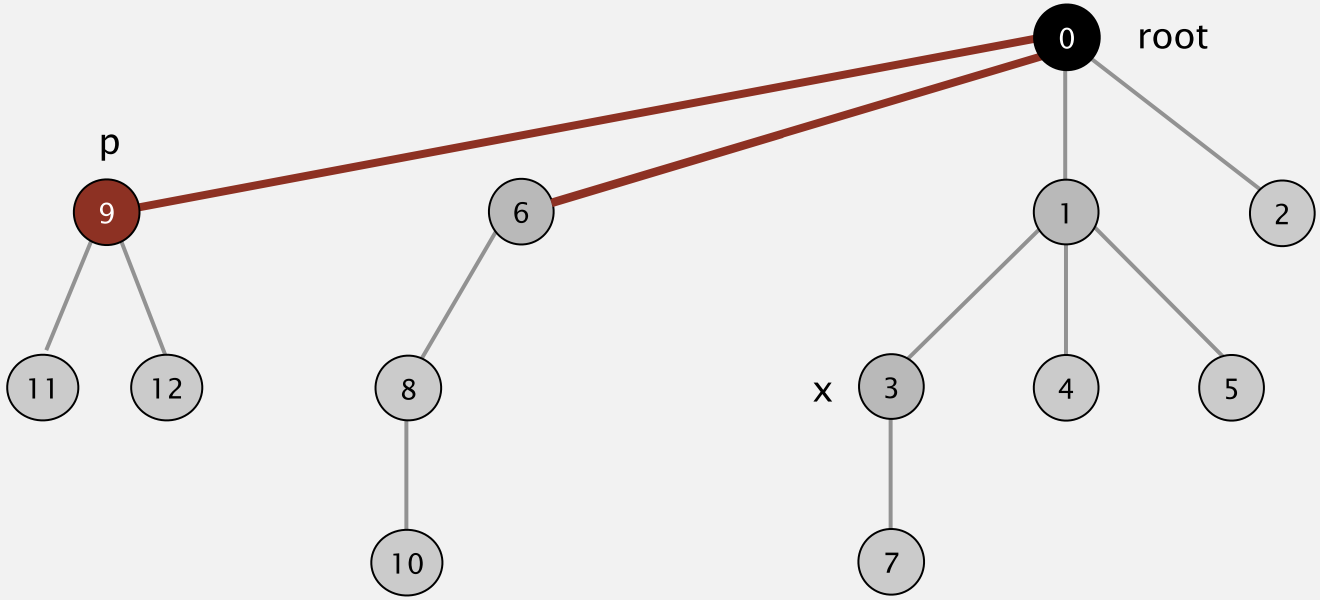
Improvement 2: path compression

Quick union with path compression. Just after computing the root of p, set the id of each examined node to point to that root.



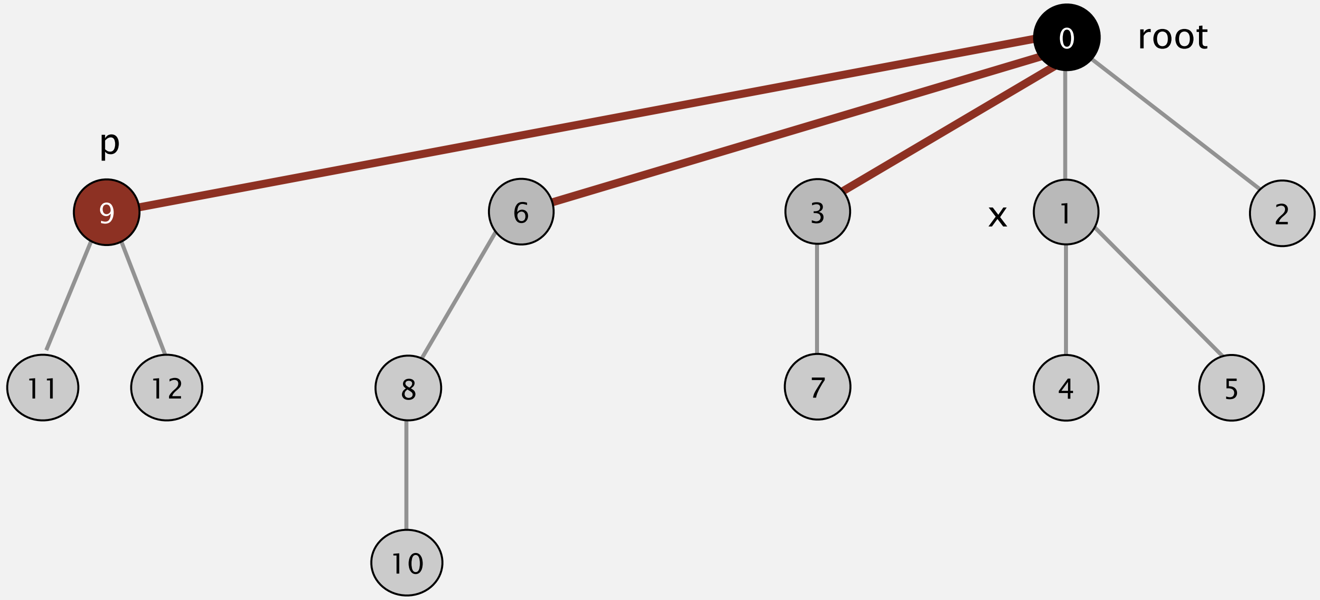
Improvement 2: path compression

Quick union with path compression. Just after computing the root of p, set the id of each examined node to point to that root.



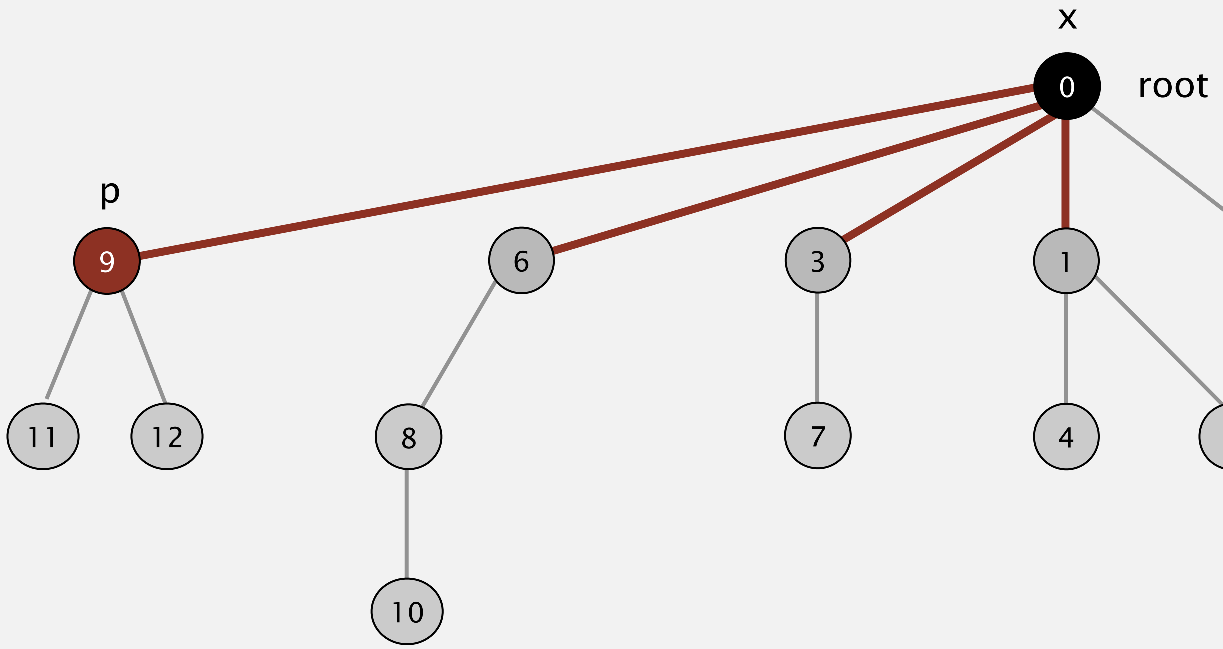
Improvement 2: path compression

Quick union with path compression. Just after computing the root of p, set the id of each examined node to point to that root.



Improvement 2: path compression

Quick union with path compression. Just after computing the root of p, set the `id[]` of each examined node to point to that root.



Path compression: Java implementation

Two-pass implementation: add second loop to `root()` to set the `id[]` of each examined node to the root.

Simpler one-pass variant: Make every other node in path point to its grandparent (thereby halving path length).

```
private int root(int i)
{
   while (i != id[i])
   {
      id[i] = id[id[i]];
      i = id[i];
   }
   return i;
}
```

only one extra line of code !

In practice. No reason not to! Keeps tree almost completely flat.

Weighted quick-union with path compression: amortized analysis

Proposition. [Hopcroft-Ulman, Tarjan] Starting from an empty data structure, any sequence of M union–find ops on N objects makes ≤ $c\,(N + M \lg^* N)$ array accesses.

- Analysis can be improved to $N + M\,\alpha(M, N)$.
- Simple algorithm with fascinating mathematics.

N	lg* N
1	0
2	1
4	2
16	3
65536	4
2^{65536}	5

iterate log function

Linear-time algorithm for M union-find ops on N objects?

- Cost within constant factor of reading in the data.
- In theory, WQUPC is not quite linear.
- In practice, WQUPC is linear.

Amazing fact. [Fredman-Saks] No linear-time algorithm exists.

in "cell-probe" model of computation

Summary

Key point. Weighted quick union (with path compression) makes it possible to solve problems that could not otherwise be addressed.

algorithm	worst-case time
quick-find	M N
quick-union	M N
weighted QU	N + M log N
QU + path compression	N + M log N
weighted QU + path compression	N + M lg* N

order of growth for M union-find operations on a set of N objects

Ex. [10^9 unions and finds with 10^9 objects]

- WQUPC reduces time from 30 years to 6 seconds.
- Supercomputer won't help much; good algorithm enables solution.

Algorithms

Robert Sedgewick | Kevin Wayne

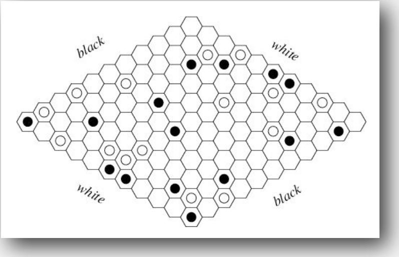
http://algs4.cs.princeton.edu

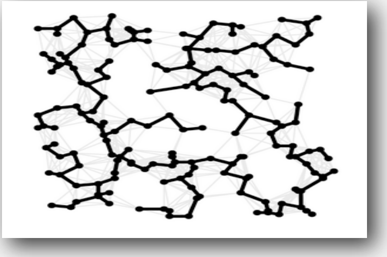
1.5 Union-Find

- *dynamic connectivity*
- *quick find*
- *quick union*
- *improvements*
- *applications*

Union-find applications

- Percolation.
- Games (Go, Hex).
- ✓ Dynamic connectivity.
- Least common ancestor.
- Equivalence of finite state automata.
- Hoshen-Kopelman algorithm in physics.
- Hinley-Milner polymorphic type inference.
- Kruskal's minimum spanning tree algorithm.
- Compiling equivalence statements in Fortran.
- Morphological attribute openings and closings.
- Matlab's `bwlabel()` function in image processing.

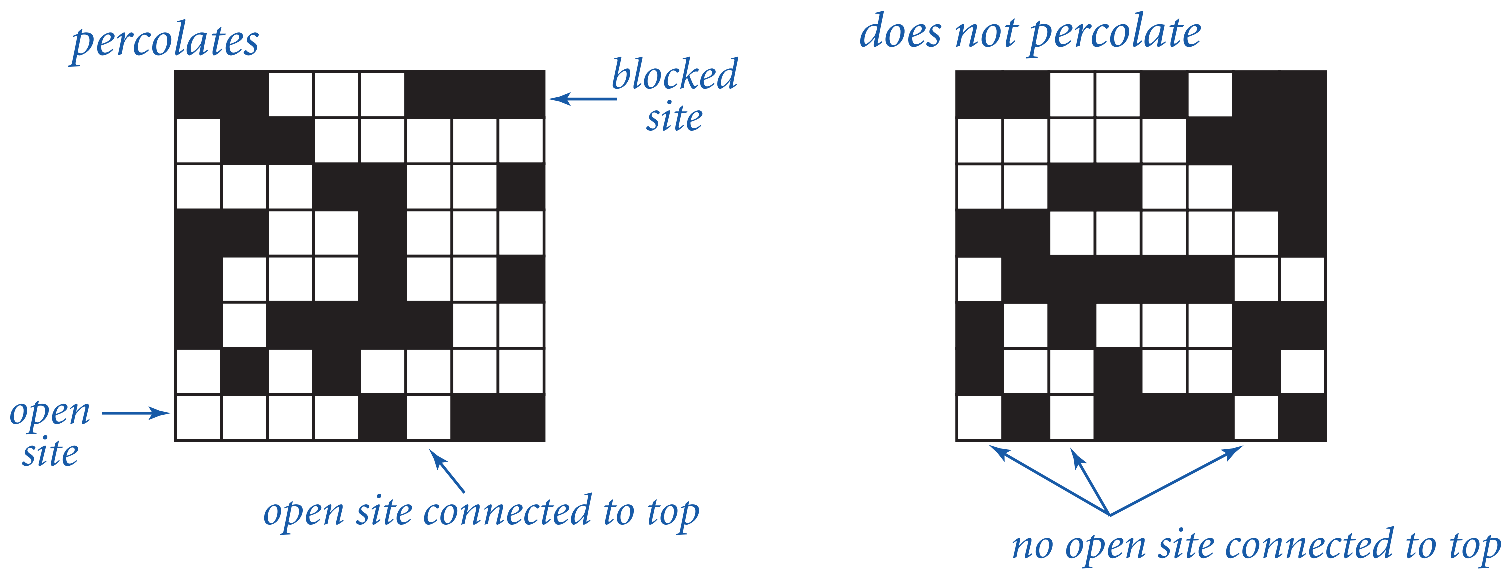




Percolation

An abstract model for many physical systems:

- N-by-N grid of sites.
- Each site is open with probability p (or blocked with probability $1 - p$).
- System percolates iff top and bottom are connected by open sites.



percolates

blocked site

open site

open site connected to top

does not percolate

no open site connected to top

N = 8

Percolation

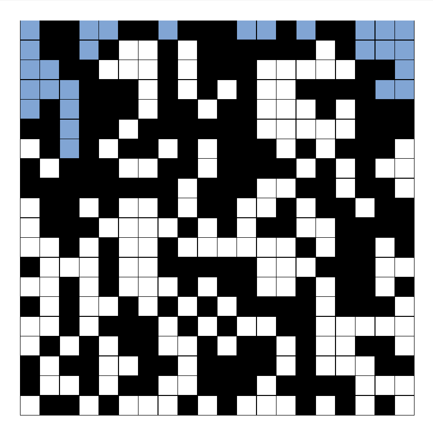
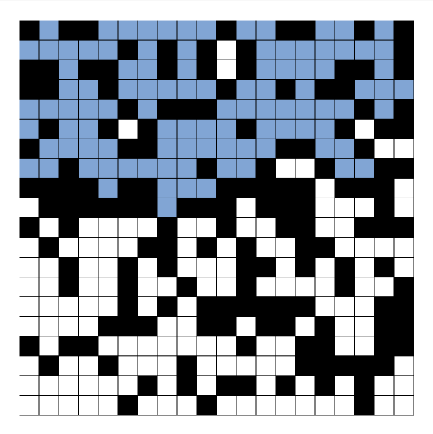
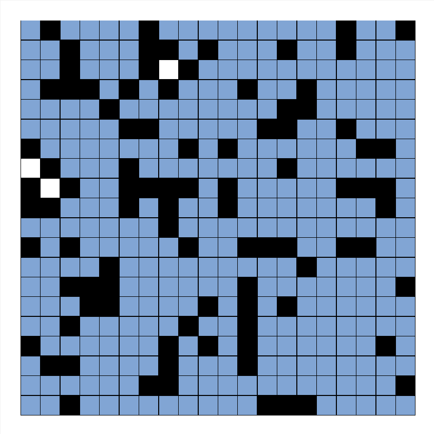
An abstract model for many physical systems:

- N-by-N grid of sites.
- Each site is open with probability p (or blocked with probability $1 - p$).
- System percolates iff top and bottom are connected by open sites.

model	system	vacant site	occupied site	percolates
electricity	material	conductor	insulated	conducts
fluid flow	material	empty	blocked	porous
social interaction	population	person	empty	communicates

Likelihood of percolation

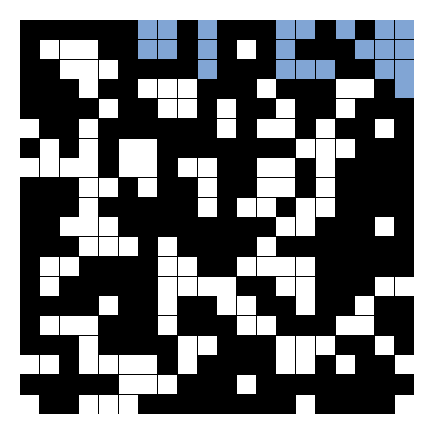
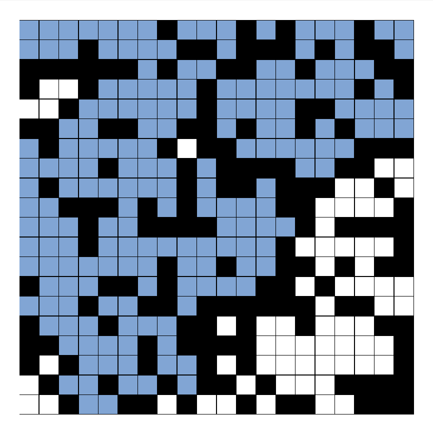
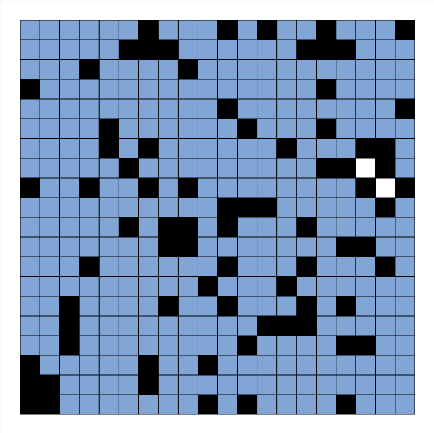
Depends on site vacancy probability p.

p low (0.4)
does not percolate

p medium (0.6)
percolates?

p high (0.8)
percolates

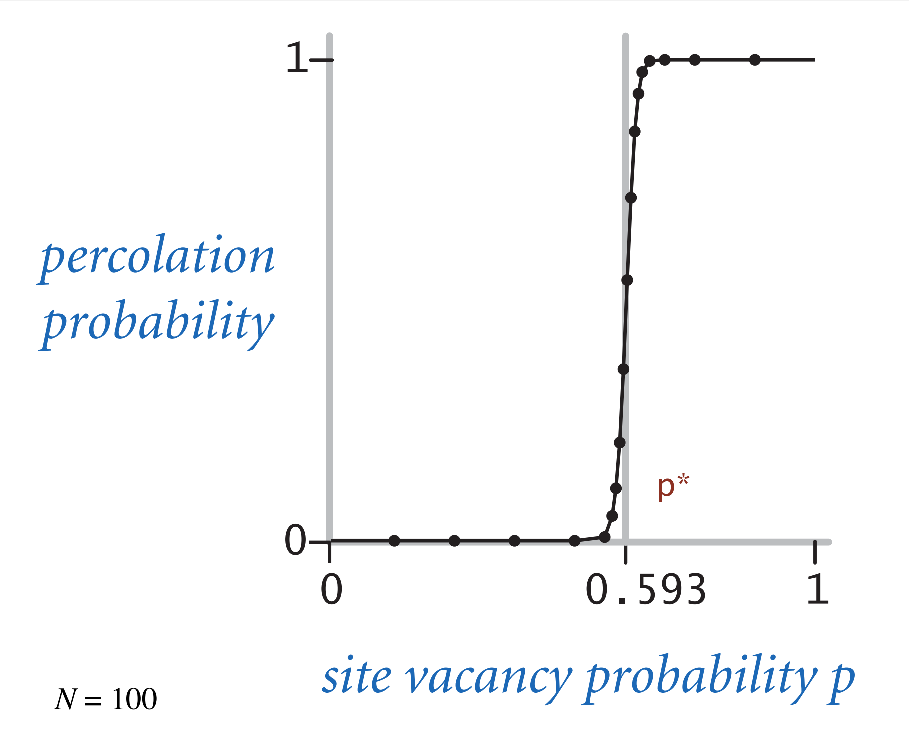




Percolation phase transition

When N is large, theory guarantees a sharp threshold p^*.

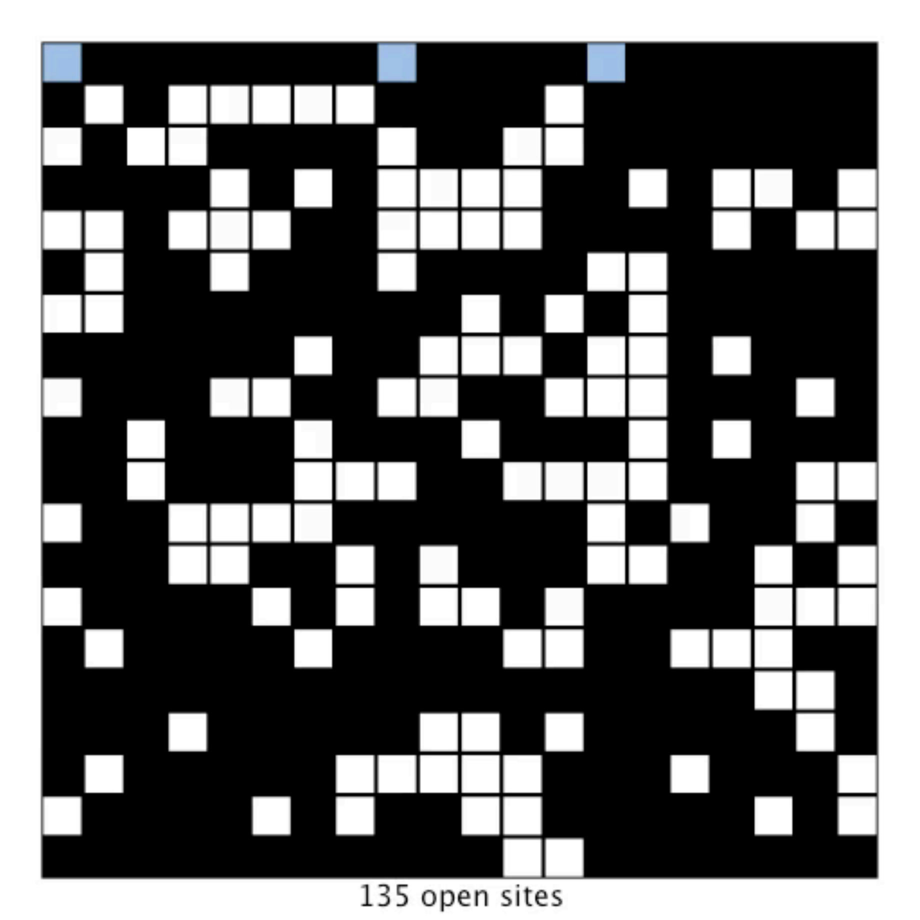
- $p > p^*$: almost certainly percolates.
- $p < p^*$: almost certainly does not percolate.

Q. What is the value of p^* ?



Monte Carlo simulation

- Initialize N-by-N whole grid to be blocked.
- Declare random sites open until top connected to bottom.
- Vacancy percentage estimates p^*.



$N = 20$

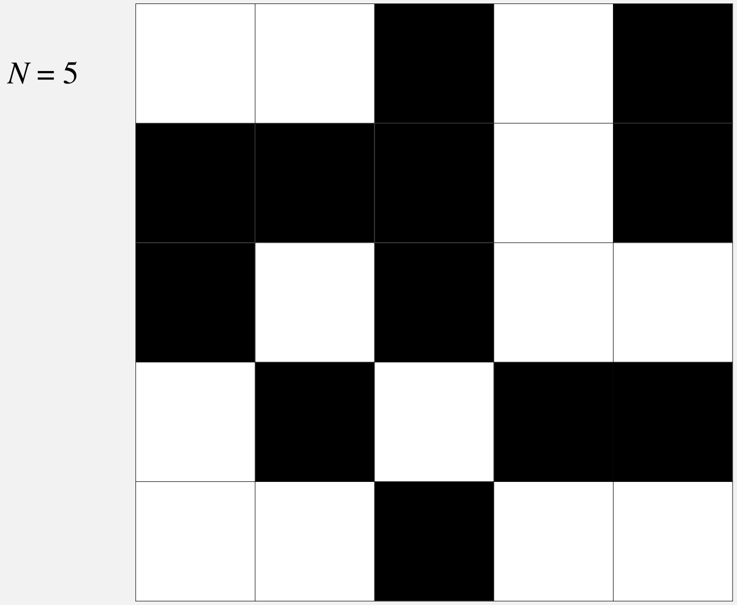
🟦 full open site (connected to top)

□ empty open site (not connected to top)

■ blocked site

Dynamic connectivity solution to estimate percolation threshold

Q. How to check whether an N-by-N system percolates?



□ open site

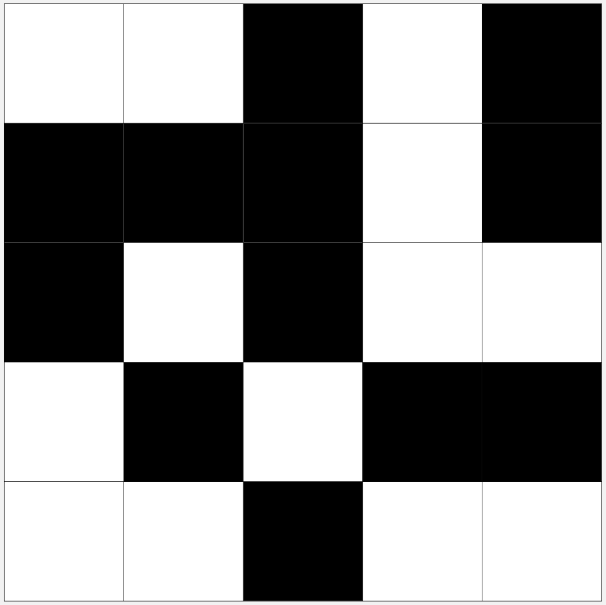
■ blocked site

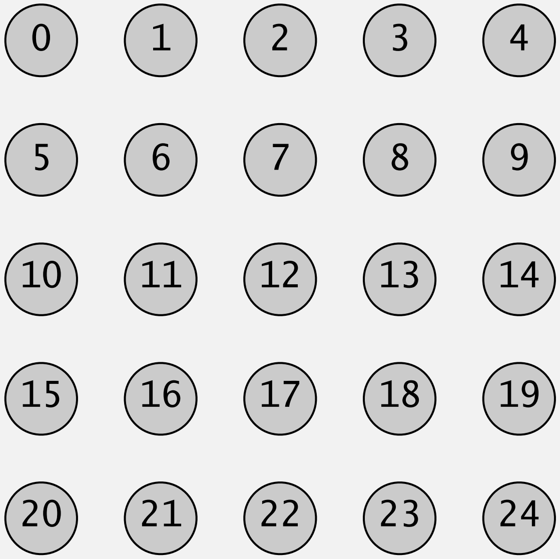
Dynamic connectivity solution to estimate percolation threshold

Q. How to check whether an N-by-N system percolates?

- Create an object for each site and name them 0 to $N^2 - 1$.

$N = 5$





□ open site

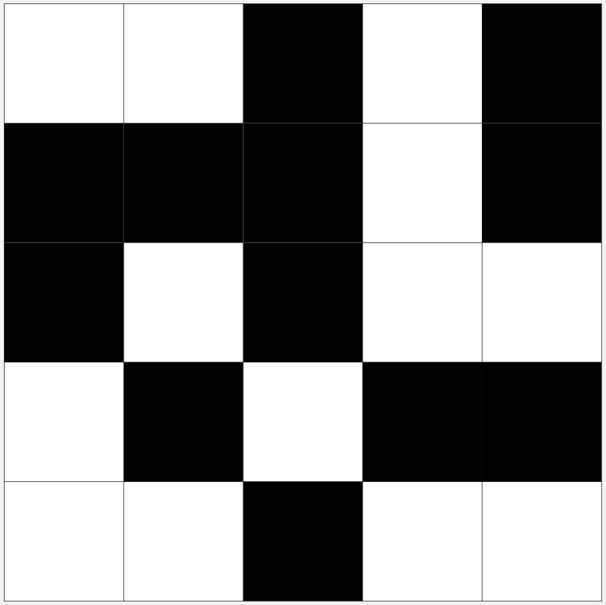
■ blocked site

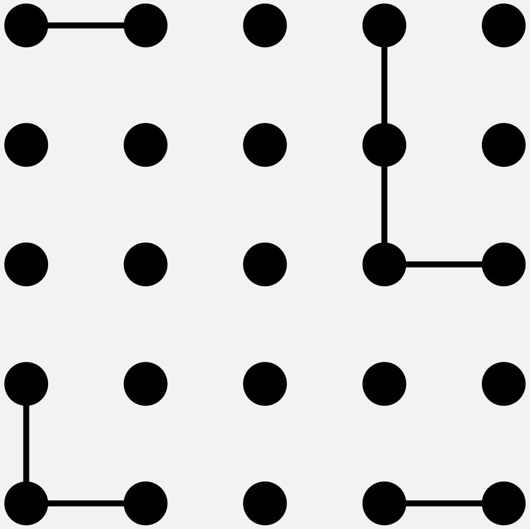
Dynamic connectivity solution to estimate percolation threshold

Q. How to check whether an N-by-N system percolates?

- Create an object for each site and name them 0 to $N^2 - 1$.
- Sites are in same component if connected by open sites.

$N = 5$





□ open site

■ blocked site

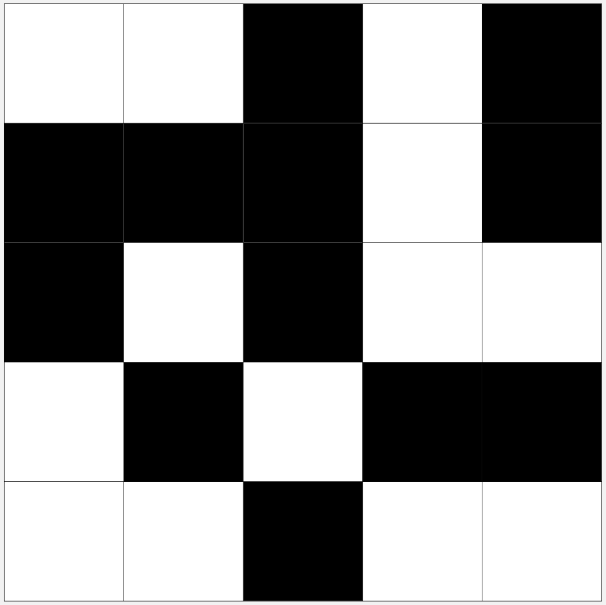
Dynamic connectivity solution to estimate percolation threshold

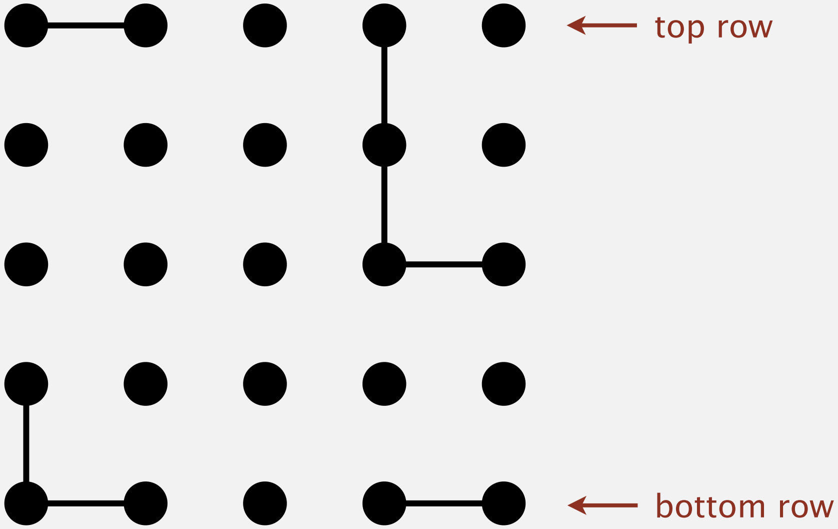
Q. How to check whether an N-by-N system percolates?

- Create an object for each site and name them 0 to $N^2 - 1$.
- Sites are in same component if connected by open sites.
- Percolates iff any site on bottom row is connected to site on top row.

brute-force algorithm: N^2 calls to connected()

$N = 5$





☐ open site

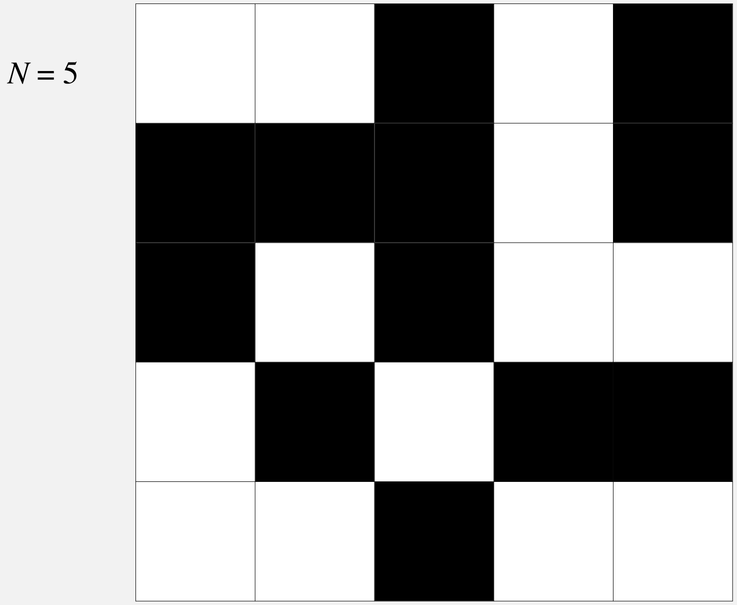
■ blocked site

Dynamic connectivity solution to estimate percolation threshold

Clever trick. Introduce 2 virtual sites (and connections to top and bottom).

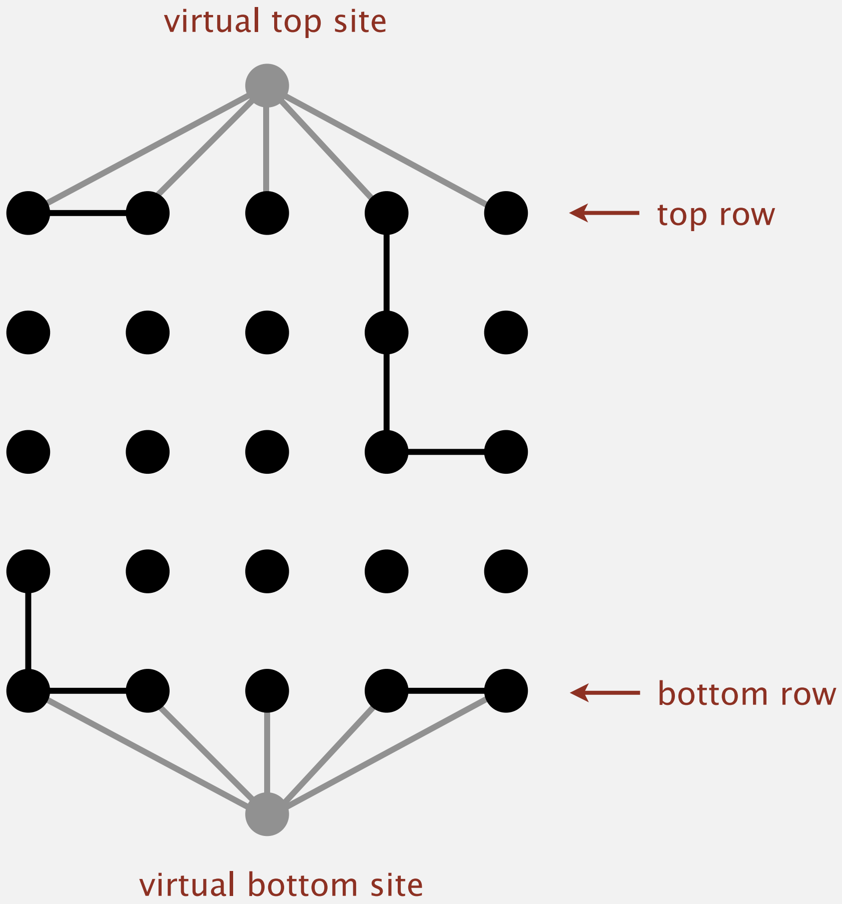
- Percolates iff virtual top site is connected to virtual bottom site.

efficient algorithm: only 1 call to connected()



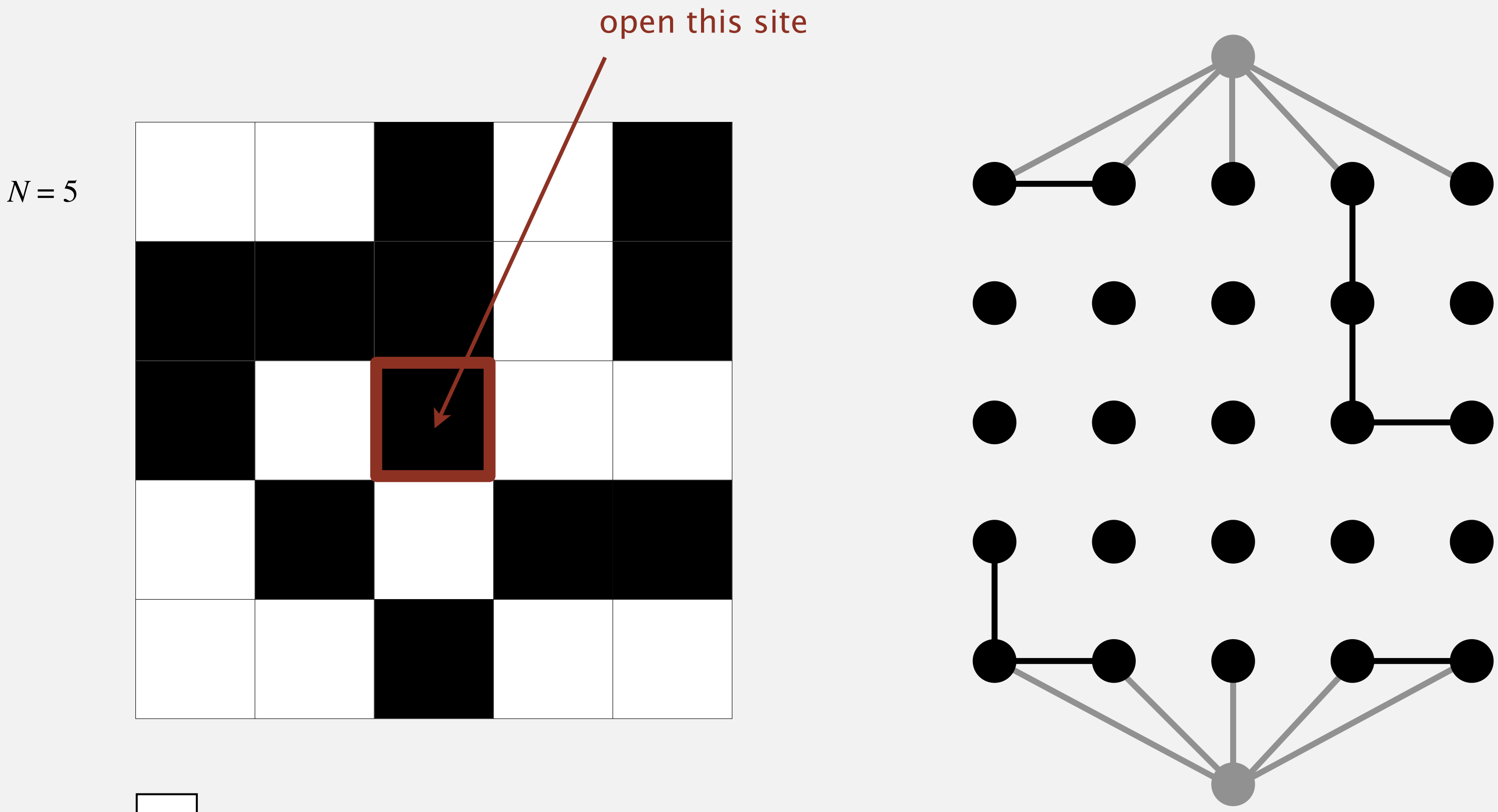
□ open site

■ blocked site



Dynamic connectivity solution to estimate percolation threshold

Q. How to model opening a new site?



□ open site

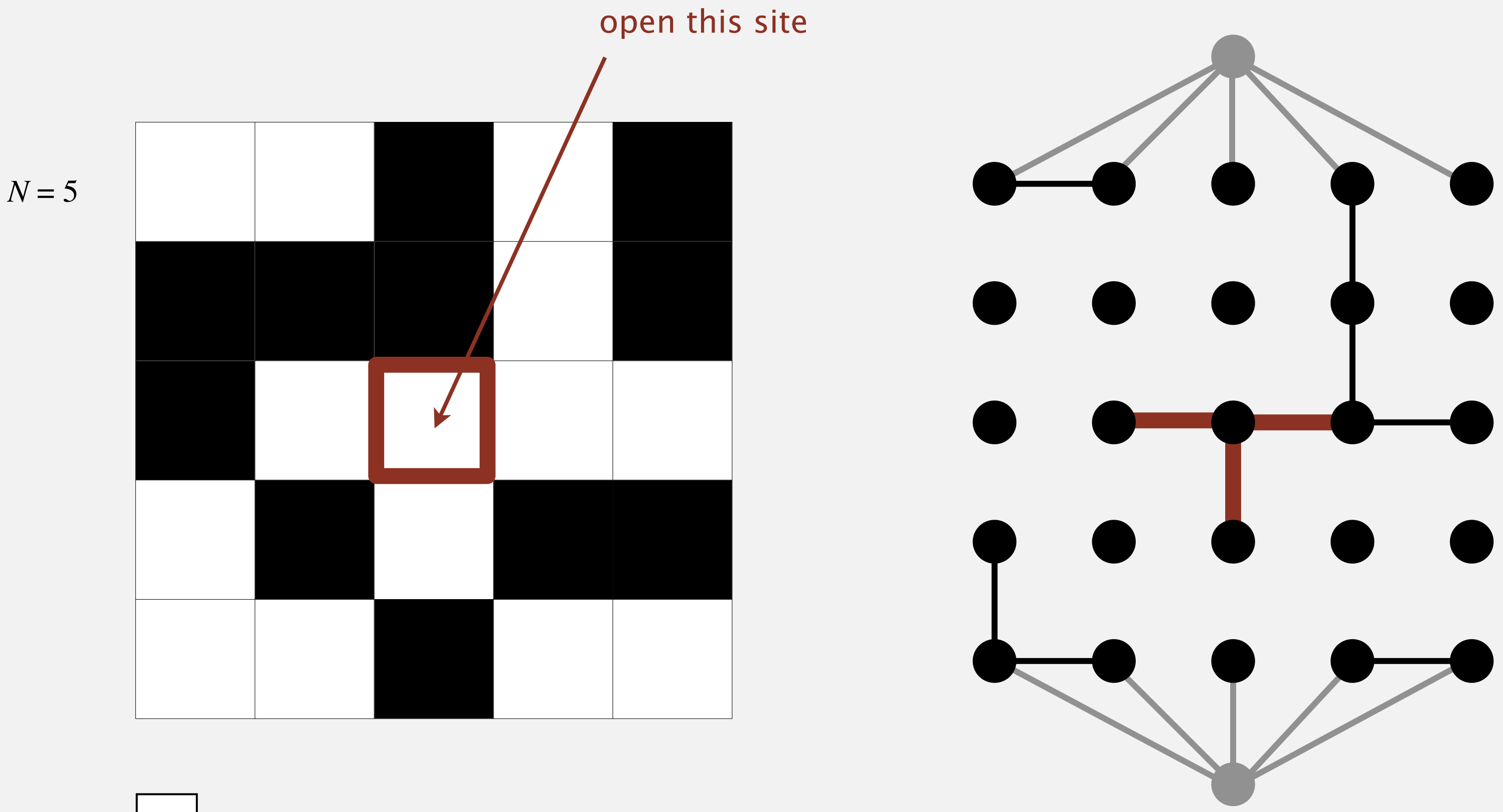
■ blocked site

Dynamic connectivity solution to estimate percolation threshold

Q. How to model opening a new site?

A. Mark new site as open; connect it to all of its adjacent open sites.

up to 4 calls to union()



□ open site

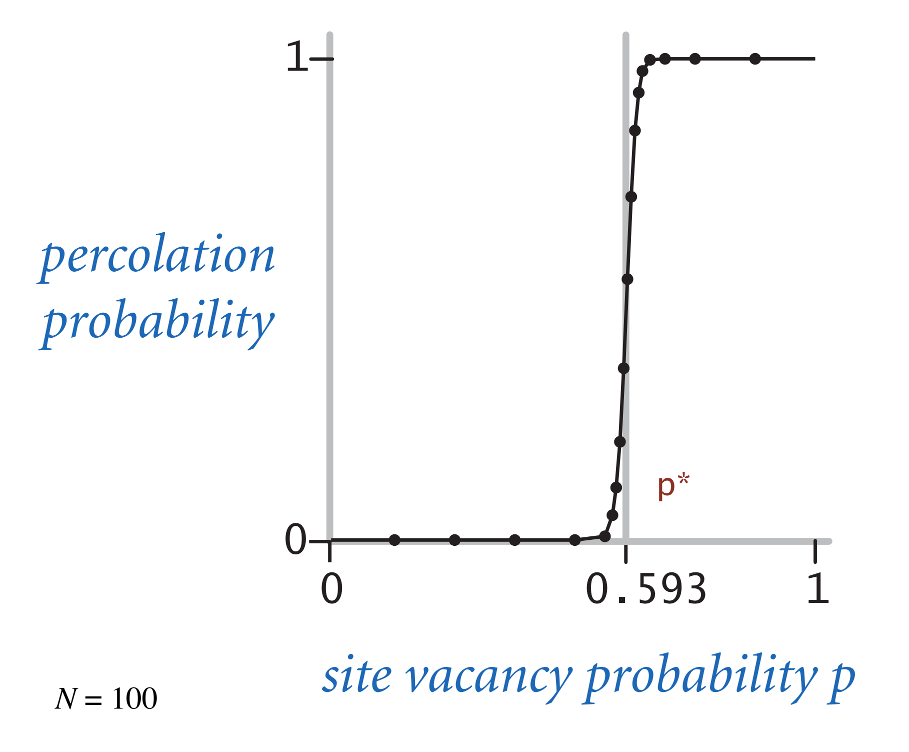
■ blocked site

Percolation threshold

Q. What is percolation threshold p^* ?

A. About 0.592746 for large square lattices.

constant known only via simulation



Fast algorithm enables accurate answer to scientific question.

Subtext of today's lecture (and this course)

Steps to developing a usable algorithm.

- Model the problem.
- Find an algorithm to solve it.
- Fast enough? Fits in memory?
- If not, figure out why.
- Find a way to address the problem.
- Iterate until satisfied.

The scientific method.

Mathematical analysis.