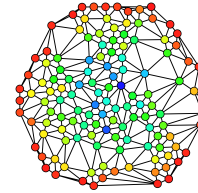


# CSE 332

## Data Abstractions

## Graphs 5: Union Find



### Kruskal's Algorithm Runtime

1

- Sort takes  $\mathcal{O}(n \lg n)$
- We don't know how UnionFind works, but if we know...
  - find is  $\mathcal{O}(\lg n)$
  - union takes  $\mathcal{O}(\lg n)$  time

The runtime is  $\mathcal{O}(|E| \lg(|E|) + |E| \lg(|V|))$

Just how does union-find work? Stay tuned!

### Disjoint Sets ADT

2

A **disjoint sets** data structure keeps track of multiple sets which do not share any elements. Here's the ADT:

#### UnionFind ADT

find(x)	Returns a number representing the set that x is in.
union(x, y)	Updates the sets so whatever sets x and y were in are now considered the same sets.

#### Example

```

1 list = [1, 2, 3, 4, 5, 6];
2 UF uf = new UF(list); // State: {1}, {2}, {3}, {4}, {5}, {6}
3 uf.find(1);           // Returns 1
4 uf.find(2);           // Returns 2
5 uf.union(1, 2);       // State: {1, 2}, {3}, {4}, {5}, {6}
6 uf.find(1);           // Returns 1
7 uf.find(2);           // Returns 1
8 uf.union(3, 5);       // State: {1, 2}, {3, 5}, {4}, {6}
9 uf.union(1, 3);       // State: {1, 2, 3, 5}, {4}, {6}
10 uf.find(3);          // Returns 1
11 uf.find(6);          // Returns 6
  
```

### Data Structure?

3

This lecture is an excuse to walk through the genesis of a data structure. We know what we want to get, and we'd like to **create (and analyze) an efficient data structure** that meets our requirements.

To define a UnionFind data structure, we need three things:

- The idea of the data structure
- An implementation of union
- An implementation of find

For the duration of the lecture, we will assume we can identify each item with a number from 1 to  $n$ .

Let's start out easy...

### Implementation 1: A List of LinkedLists

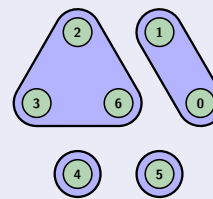
4

#### Data Structure

**Type:** List<LinkedList<Integer>>

**Idea:** A mapping from id  $\rightarrow$  a list of ids in the same set

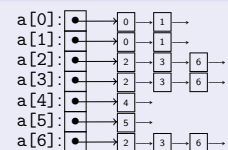
#### Pictorial View



```

find(x)
1 find(x) {
2   return a[x].front;
3 }
  
```

#### Data Structure



```

union(x, y)
1 union(x, y) {
2   ...
3 }
  
```

## Implementation 1.5: A List of LinkedLists

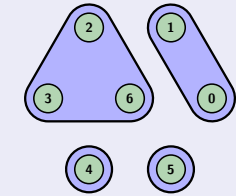
5

### Data Structure

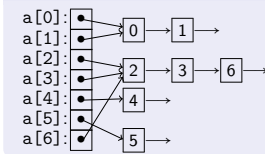
**Type:** List<LinkedList<Integer>>

**Idea:** A mapping from **id** → a list of **ids** in the same set

### Pictorial View



### Data Structure



### union(x, y)

```

1 union(x, y) {
2   curr = a[x].head;
3   a[y].tail.next = curr;
4   while (curr != null && curr.next != null) {
5     a[curr.data] = a[y].head
6     curr = curr.next;
7   }
8 }

```

### find(x)

```

1 find(x) {
2   return a[x].front;
3 }

```

## Implementation 1.5: A List of LinkedLists

6

### Data Structure

**Type:** List<LinkedList<Integer>>

**Idea:** A mapping from **id** → a list of **ids** in the same set

### find(x)

```

1 find(x) {
2   return a[x].front;
3 }

```

### union(x, y)

```

1 union(x, y) {
2   curr = a[x].head;
3   a[y].tail.next = curr;
4   while (curr != null && curr.next != null) {
5     a[curr.data] = a[y].head
6     curr = curr.next;
7   }
8 }

```

### Asymptotic Analysis

- find(x) is  $\mathcal{O}(1)$
- union(x, y) is  $\mathcal{O}(a[x].length)$

### Amortized Analysis

Consider any  $m$  find/union operations. The **worst** case is going to be that all the operations are all unions, but which unions?

Always union(LARGEST, y), because we have to traverse the first one.

This ends up being  $1 + 2 + \dots + n - 1 = \frac{(n-1)n}{2}$ . This is a total of  $n - 1$

## Implementation 2: A List of LinkedLists Unioned-By-Weight

7

### Data Structure

**Type:** List<LinkedList<Integer>>

**Idea:** A mapping from **id** → a list of **ids** in the same set

### OLD union(x, y)

```

1 union(x, y) {
2   curr = a[x].head;
3   a[y].tail.next = curr;
4   while (curr != null && curr.next != null) {
5     a[curr.data] = a[y].head;
6     curr = curr.next;
7   }
8 }

```

### NEW union(x, y)

```

1 union(x, y) {
2   if (a[x].length > a[y].length) {
3     x, y = swap(x, y);
4   }
5   curr = a[x].head;
6   a[y].tail.next = curr;
7   while (curr != null && curr.next != null) {
8     a[curr.data] = a[y].head;
9     curr = curr.next;
10  }
11 }

```

### Asymptotic Analysis

- find(x) is  $\mathcal{O}(1)$
- union(x, y) is  $\mathcal{O}(\min(a[x].length, a[y].length))$

### Amortized Analysis

Consider any  $m$  find/union operations. The **worst** case is going to be that all the operations are all unions, but which unions?

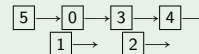
## Implementation 3: IMPLICIT Lists Unioned-By-Size

8

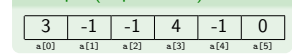
We started with a **list of linked lists**. Then, we realized that we could use **references to the same linked list** to save memory.

We can do even better. The idea is to use an "implicit list".

### Example (Explicit List)



### Example (Implicit List)



If you've already taken CSE 351, you've seen this idea already! When implementing malloc, you store a **free list**. You can save a lot of memory (which in malloc is important...) by using the unused **data fields** to store the **pointers**.

### Using An Implicit List

We need to store:

- pointers to get to the canonical member
- the size of the set

## Implementation 3: IMPLICIT Lists Unioned-By-Size

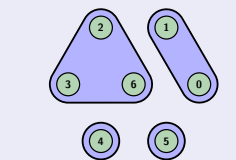
9

### Data Structure

**Type:** An array

**Idea:** Each index has either the value of the "next" thing in its set or a negative number representing the size of the set

### Pictorial View



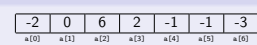
### Implementation

```

1 init(x) { a[x] = -1 }
2 find(x) {
3   while (a[x] >= 0) {
4     x = a[x]
5   }
6   return x
7 }
8
9 size(x) { return -a[find(x)] }
10
11 union(x, y) {
12   if (size(x) > size(y)) {
13     x, y = swap(x, y)
14   }
15   // Now, we have: size(x) <= size(y)
16   a[find(x)] = find(y)
17   // Update the size
18   a[find(y)] = size(x) + size(y)
19 }
20 }
21 }

```

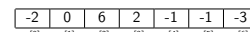
### Data Structure



- "Non-canonicals" store "pointers"
- "Canonicals" store -size

## Analyzing Implementation 3

10



### Implementation

```

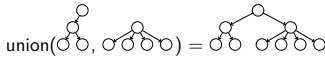
1 init(x) { a[x] = -1 }
2 find(x) {
3   while (a[x] >= 0) {
4     x = a[x]
5   }
6   return x
7 }
8
9 size(x) { return -a[find(x)] }
10
11 union(x, y) {
12   if (size(x) > size(y)) {
13     x, y = swap(x, y)
14   }
15   // Now, we have: size(x) <= size(y)
16   a[find(x)] = find(y)
17   // Update the size
18   a[find(y)] = size(x) + size(y)
19 }
20 }
21 }

```

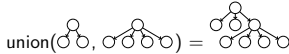
- Assume we only call each size/find once.
- Then, union(x, y) ∈  $\mathcal{O}(\text{find}(x) + \text{find}(y))$ .
- So, we only need analyze find(x).
- We claim that find(x) ∈  $\mathcal{O}(\lg n)$ .
- To prove this, we will show the **height** of the tree resulting from some number of unions is  $\mathcal{O}(\lg n)$
- (Sound familiar?)

We claim:

- If  $x, y$  are different heights,  
 $\text{height}(\text{union}(x, y)) = \max(\text{height}(x), \text{height}(y))$



- If  $x, y$  are the same height,  
 $\text{height}(\text{union}(x, y)) = 1 + \max(\text{height}(x), \text{height}(y))$



Let  $f(h)$  be the number of (distinct) nodes necessary to make a tree of height  $h$ . Then,  $f(1) = 1$  and  $f(h) = 2f(h-1)$ . So,  $f(h) = 2^{h-1}$ .

Then, what is the largest tree we can make with  $n$  nodes?

$$n = 2^{h-1} \rightarrow \lg(n) = h - 1$$

So,  $h \in \mathcal{O}(\lg n)$ .

(Notice this is nearly identical to the proof for AVL Trees, except with a slightly different recurrence.)

OLD find(x)

```
1 find(x) {
2   while(a[x] >= 0) {
3     x = a[x]
4   }
5   return x
6 }
```

NEW find(x)

```
1 find(x) {
2   if (a[x] < 0) {
3     return x
4   }
5   a[x] = find(a[x])
6   return a[x]
7 }
```

In Words: Once we've found a node... save it.

Amortized Analysis of  $m$  find Operations?

Consider what we know:

- We know the worst case height of a tree is  $\lg(n)$ .
- We know it's difficult to make a tree of large height.
- We know that as soon as we access a path in a tree, it flattens the whole path

This feels like it should be better than  $\lg(n)$ , and it is.

We can use facts to show this, but its outside the scope of this lecture. Instead, we'll just talk about two bounds.

Upper Bound 1:  $\text{find}(x)$  is amortized  $\mathcal{O}(\lg^*(n))$

Let  $2\text{STACK}(n) = 2^{2^{\dots^2}}$

- $2\text{STACK}(0) = 1$
- $2\text{STACK}(1) = 2$
- $2\text{STACK}(2) = 2^2 = 4$
- $2\text{STACK}(3) = 2^{2^2} = 16$
- $2\text{STACK}(4) = 2^{2^{2^2}} = 2^{16} = 65536$
- $2\text{STACK}(5) = 2^{65536}$

$2^{65536} = 2003529930406846464979072351560255750447825475569751419 \dots$   
 2650169737108940595563114530895061308809333481010382343429072  
 6318182294938211881266886950636476154702916504187191635158796  
 6347219442930927982084309104855990570159318959639524863372367  
 2030029169695921561087649488892540908059114570376752085002066  
 7156370236612635974714480711177481588091413574272096719015183  
 6282560618091458852699826141425030123391108273603843767876449  
 0432059603791244909057075603140350761625624760318637931264847  
 0374378295497561377098160461441330869211810248595915238019533  
 1030292162800160568670105651646750568038741529463842244845292  
 537361442533614373290883037946012747249584148649159306472520  
 151569392262818069165079638106413227530726714399815850881129  
 262890113423778270556742108007006528396322155077831214288551  
 6755540733451072131124273995629827197691500548839052238043570  
 4584819795639315785351001899200002414196370681355984046403947  
 2194016069517690156119726982337890017641517190051133466306898  
 1402193834814354263873065395529696913880241581618595611006403  
 6211979610185953480278716720012260464249238511139340046435162  
 3867567078745259464670903886547743483217897012764455529409092  
 0219595857516229733335761595523948852975799540284719435299135

4376370598692891375715374000198639433246489005254310662966916  
 5243419174691389632476560289415199775477703138064781342309596  
 1909606545913008901888875880847336259560654448885014473357060  
 5881709016210849971452956834406197969056546981363116205357936  
 9791403236328496233046421066136200220175787851857409162050489  
 7117818204001872829399434461862243280098373237649318147898481  
 1945271300744022076568091037620399920349202390662626449190916  
 7985461515778839060397720759279378852241294301017458086862263  
 3692847258514030396155585643303854506886522131148136384083847  
 7826379045960718687672850976347127198889068047824323039471865  
 0525660978150729861141430305816927924971409161059417185352275  
 8875044775922183011587807019755357222414000195481020056617735  
 8978149953232520858975346354700778669040642901676380816174055  
 0405117670093673202804549339027992491867306539931640720492238  
 4748152806191669009338057321208163507076343516698696250209690  
 2316285935007187419057916124153689751480826190484794657173660  
 1005892476655445840838334790544144817684255327207315586349347  
 6051374197795251903650321980201087647383686825310251833775339  
 0886142618480037400808223810407646887847164755294532694766170  
 0424461063311238021134588694532200116564076327023074292426051

6340696503084422585596703927186946115851379338647569974856867  
 0079823960604393478850861649260304945061743412365828352144806  
 7266768418070837548622114082365798029612000274413244384324023  
 3125740354501935242877643088023285085588608996277445816468085  
 7875115807014743763867976955049991643998284357290415378143438  
 8473034842619033888414940313661398542576355771053355802066221  
 855770600825512888933222643628198483861323957067619140963853  
 383237434375883085923372228464287996245605476932428998432652  
 6773783731732880632107532112386806046747084280511664887090847  
 702912081611049125559832236624486855665140268464120969498259  
 0565519216188104341226838996283071654868525536914850299539675  
 5039549383718534059000961874894739928804324963731657538036735  
 8671017578399481847179849824694806053208199606618343401247609  
 6639519778021441199752546704080608499344178256285092726523709  
 8986515394621930046073645079262129759176982938923670151709920  
 915315678144397912484757062378046000991829332130688057004659  
 1458387208088016887445835557926258465124763087148566313528934  
 166117490617526671492672176128330845273936469244828925713888  
 7783905630048248379983969202922215486145902373478222628252163  
 9957440801727144146179559226175083889020074169926238300282286



