Samuel Osa-Agbontaen

CS120: Intro. to Algorithms and their Limitations Hesterberg & Vadhan

Problem Set 0

Harvard SEAS - Fall 2022 *Due: Wed 2023-09-13 (11:59PM)*

The purpose of this problem set is to reactivate your skills in proofs and programming from CS20 and CS32/CS50. For those of you who haven't taken one or both those courses, the problem set can also help you assess whether you have acquired sufficient skills to enter CS120 in other ways and can fill in any missing gaps through self-study. Even for students with all of the recommended background, this problem set may still require a significant amount of thought and effort, so do not be discouraged if that is the case and do take advantage of the staff support in section and office hours.

For those of you who are wondering whether you should wait and take CS20 before taking CS120, we encourage you to also complete the CS20 Placement Self-Assessment. Some problems there that are of particular relevance to CS120 and are complementary to what is covered below are Problems 2 (counting), 4 (comparing growth rates), 9 (quantificational logic), and 12 (graph theory).

Written answers must be submitted in pdf format on Gradescope. Although LaTeX is not required, it is strongly encouraged. You may handwrite solutions so long as they are fully legible. The `ps0` directory, which contains your code for problems 1a and 1c, must be submitted separately to an autograder on Gradescope. Be sure to pull the starter code from the cs120 GitHub repository.

1. (Binary Trees) In the cs120 GitHub repository, we have given you a Python implementation of a binary tree data structure, as well as a collection of test trees built using this data structure. We specify a binary tree by giving a pointer to its *root*, which is a special *vertex* (a.k.a. *node*), and giving every vertex pointers to its *children* vertices and its *parent* vertex as well as an identifying *key*:

```
class BinaryTree:
    def __init__(self, root):
        self.root: BTvertex = root

class BTvertex:
    def __init__(self, key):
        self.parent: BTvertex = None
        self.left: BTvertex = None
        self.right: BTvertex = None
        self.key: int = key
        self.size: int = None
```

In CS50, the concept of a Python `class` was not covered. Here, with `BinaryTree` and `BTvertex`, we are using them in the same way as a `struct` in C. An object `v` of the `BTvertex` class contains five attributes, which we list with the type of the object we expect to be named by each attribute (using the Python type annotation syntax). These attributes can be accessed as `v.parent`, `v.left`, `v.right`, `v.key`, and `v.size`. For example, `v.left.key` is the key associated with `v`'s left child. An object of the `BinaryTree` class contains only one attribute, which is the `BTvertex` object that is the root of our binary tree. You can create a `BinaryTree` object as follows:

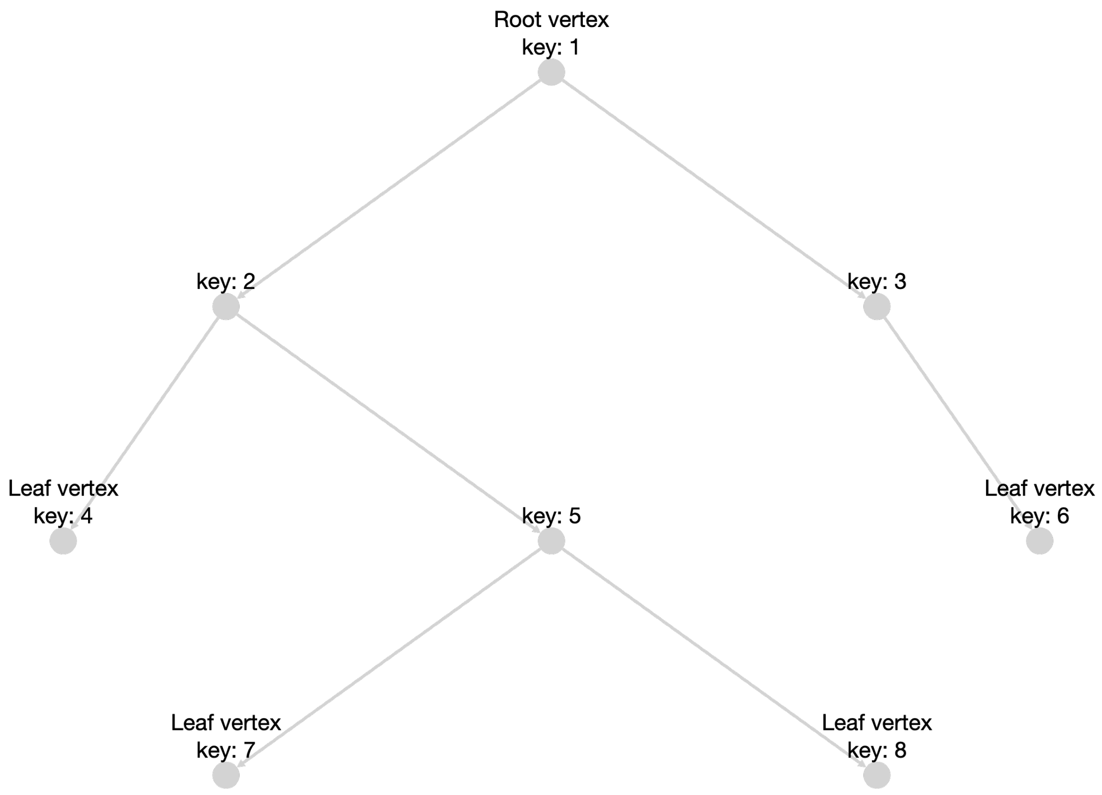
```
root = BTvertex(120)
tree = BinaryTree(root)
tree.root.left = BTvertex(121)
tree.root.right = BTvertex(124)
```

You can then print attributes of the newly created `BinaryTree` object:

```
print(tree.root.key)
>> 120
print(tree.root.left.key)
>> 121
```

Classes are more general than structs because they can also have private attributes and methods that operate on the attributes, allowing for object-oriented programming. However, you won't need that generality in this problem set.

Here is an instance `T` of `BinaryTree`:



A `BinaryTree` `T` contains only a pointer to its root vertex, `T.root`, which is required to satisfy `T.root.parent==None`. In the above example, the root is the vertex with key 1 (i.e. `T.root.key==1`). A binary tree vertex `v` can have zero, one, or two children, determined by which of `v.left` and `v.right` are equal to `None`. In the above example, the vertex `v` with key 3 has `v.left==None` but `v.right` is the vertex with key 6. A *leaf* is a vertex with zero children, i.e. `v.left==v.right==None`.

A vertex `w` is *descendent* of a vertex `v` if there is a sequence of vertices $\mathtt{v}_0, \mathtt{v}_1, \ldots, \mathtt{v}_k$, $k \in \mathbb{N}$ such that $\mathtt{v}_0 = \mathtt{v}$, $\mathtt{v}_k = \mathtt{w}$, and $\mathtt{v}_i \in \{\mathtt{v}_{i-1}.\mathtt{left}, \mathtt{v}_{i-1}.\mathtt{right}\}$ for $i = 1, \ldots, k$.[1] In the above example, the vertex with key 5 is a descendent of the root (with a path of length 2), but is not a descendent of the vertex with key 3. The sequence $\mathtt{v}_0, \mathtt{v}_1, \ldots, \mathtt{v}_k$ is called a *path* from `v` to `w` and k is the *distance* from `v` to `w`. Taking $k = 0$, we see that `v` is a descendent of itself.

The *vertex set* of a binary tree `T` consists of all of the descendents of `T.root`. The *size* of `T` is its number of vertices. The *height* of `T` is the largest distance from the root to a leaf. The above example has size 8 and height 3.

Given any vertex `v` in a tree, the *subtree* rooted at `v` consists of all of `v`'s descendents. Note that we can remove a subtree and turn it into a new tree `S` by setting `S.root=v` and `v.parent=None`.

For now, the `key` attribute serves to distinguish vertices from each other in our tests and help illustrate what the algorithms are doing. The `BTvertex` class also has a `size` attribute, which is initialized to `None` in all of the test instances; it will be filled in by the program you write in Part 1a.

An instance `T` `BinaryTree` is *valid* if it satisfies the following constraints:

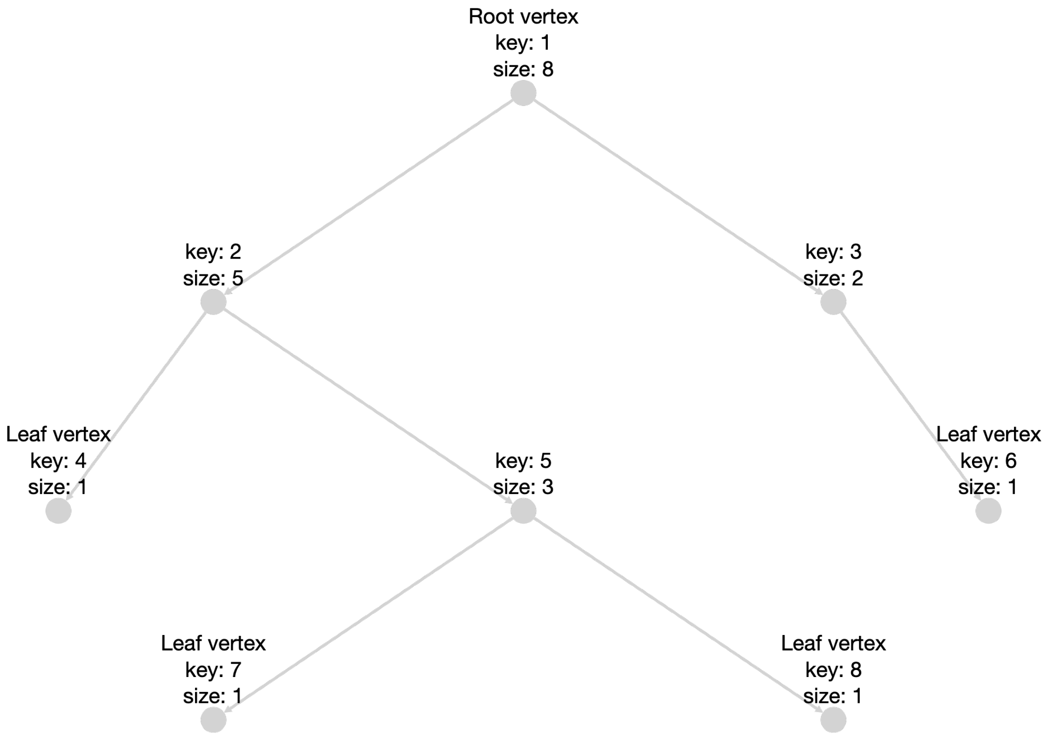
- `T.root.parent==None`
- `T` has finitely many vertices.
- No two vertices `v`, `w` of `T` share a child, i.e. $\{\mathtt{v.left}, \mathtt{v.right}\} \cap \{\mathtt{w.left}, \mathtt{w.right}\} = \emptyset$.

All of the test instances we provide are valid, and furthermore have the property that all of the vertices have distinct keys (which is something we often want, but not always).

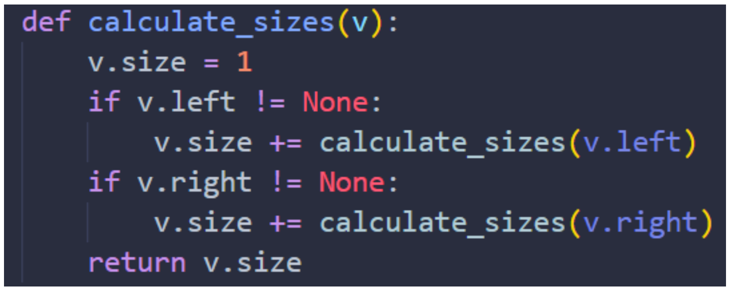
(a) (recursive programming) Write a recursive program `calculate_sizes` that given a vertex `v` of a binary tree `T`, calculates the sizes of all of the subtrees rooted at descendents of `v`. After running your program on `T.root`, every vertex `v` in `T` should have `v.size` set to the size of the subtree rooted at `v`. (Recall that the size attributes are initialized to `None`.) We call the resulting tree a *size-augmented* tree.

For example, if `T` is the tree shown above, then calling `calculate_sizes(T.root)` should modify `T` to be the following size-augmented tree:

[1] $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, 3, \ldots\}$. Since we are computer scientists, we start counting at 0.



Your program should run in time $O(n)$ when given the root of a tree with n vertices. In a sentence or two, informally justify why your program has such a runtime.

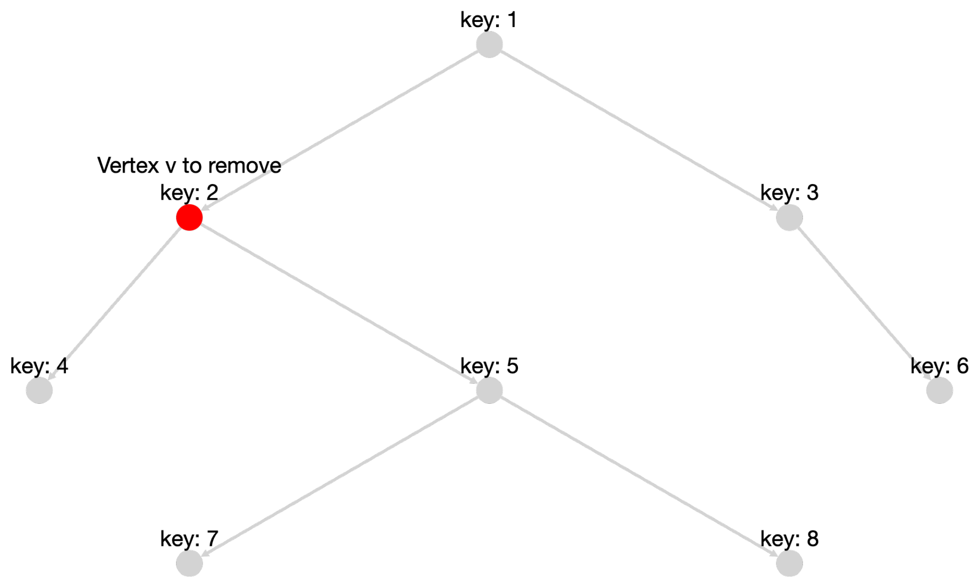


```
def calculate_sizes(v):
    v.size = 1
    if v.left != None:
        v.size += calculate_sizes(v.left)
    if v.right != None:
        v.size += calculate_sizes(v.right)
    return v.size
```

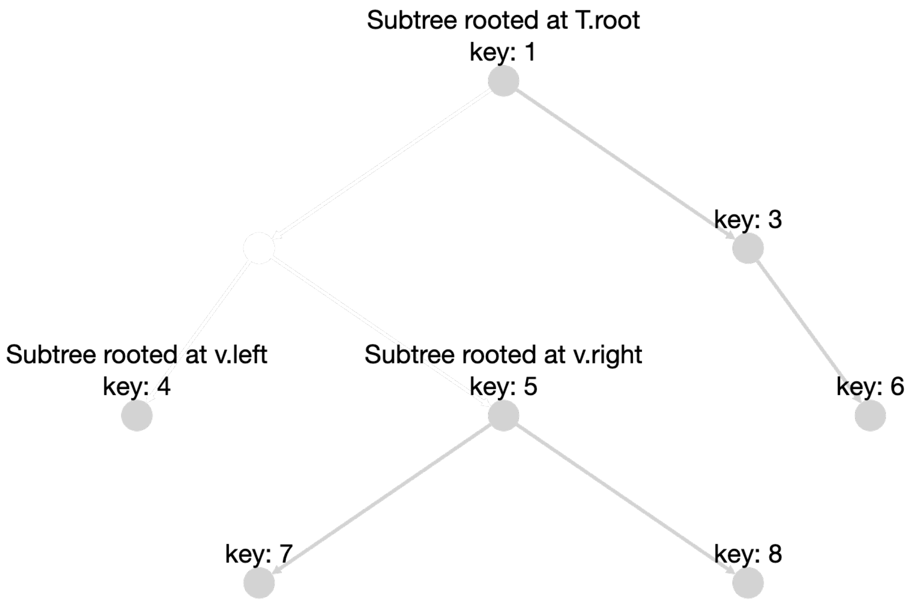
This runs in $O(n)$ time because we only visit each node once, so the runtime will have a linear relationship with the input size (number of nodes in the tree). The recursion stack does not contribute to the time complexity, but rather, the space complexity.

(b) (proof warmup) Removing a vertex `v` from a tree `T` yields up to three disjoint trees: the subtree rooted at `v.left` (unless `v.left==None`), the subtree rooted at `v.right` (unless `v.right==None`), and a tree rooted at `T.root` consisting of all non-descendants of `v` (unless `T.root==v`). For example:

Before:



After:



Suppose that for a vertex `v`, the largest subtree created by removing `v` contains a neighbor (i.e. child or parent) `v`* of `v` and a total of $\phi($`v`$)$ vertices (including `v`*). If we remove `v`* and not `v`, we create at most two subtrees which don't contain `v`: prove that these subtrees each contain fewer than $\phi($`v`$)$ vertices.

Given that removing v generates the largest subtree (A) with size $\phi(v)$, and $v*$ is v's neighbor and serves as the root node for A, we know that removing $v*$ will create up to two child subtrees that do not contrain v. Since $v*$ is the root of A, we know that A is entirely made up of $v*$ and its children subtrees B and C. Even though we do not know the size of B and C, we know that $|B| + |C| + 1 = |A|$, (where the "1" where represents the $v*$ node), which becomes $|B| + |C| + 1 = \phi(v)$, also written as $|B| + |C| = \phi(v) - 1$. We know that $\phi(v)$ is positive, and for this equality to hold, $|B|$ and $|C|$ must also be positive, and for them to both be addends that add up to $\phi(v) - 1$, they must be smaller than $\phi(v)$.

(c) (proofs by contradiction) Prove that in every tree `T` of size n, there exists a vertex `v` such that removing `v` from `T` results in disjoint trees that all have size at most $n/2$.

You may prove this however you like, but a recommended approach is to extend Part 1b and show that if the largest subtree created by removing a vertex `v` contains a neighbor `v`* of `v` and contains $\phi(\mathtt{v}) > n/2$ vertices, then if we remove `v`* and not `v`, the subtree *containing* `v` contains at most $n/2$ vertices. Then choose `v` to be the vertex for which $\phi(\mathtt{v})$, the size of the largest tree created by removing `v`, is smallest.
For purposes of contradiction, we will suppose that no vertex v exists in T with potential $< n/2$. In other words, for all v, there is always going to be a subtree where $\phi(v) > n/2$. This is the same as saying, there exists a v' which yields the lowest potential (ϕ is minimized). Since this minimizes potential, then if we can show that this case has potential $> n/2$, then it follows that all other v will be yield a potential greater than $n/2$. However, this supposition falls apart if we can find a vertex that yields an even lower potential than v'. We can do this by considering our result from part 1b, which shows that there exists vertices that can yield lower potential than v. Specifically, if we remove a neighbor($v*$) of v', the subtrees generated that do not include v' will be smaller than $\phi(v')$ Therefore, the supposition is actually false, meaning the original statement is true.

(d) (from proofs to algorithms) Turn your proof from Part 1c into a Python program that given a root vertex `r` of a *size-augmented* tree `T` with n vertices finds a vertex `v` with $\phi(\mathtt{v}) \leq n/2$. Your program should run in time $O(h)$ on all size-augmented trees of height h; again informally justify why your program has such a runtime. (Hint: try to repeatedly reduce the potential function by moving to children. Why don't we need to try moving to parents as in the previous proof?) We only traverse through each level of the node once. We don't need to move to the parents because we can calculate the size of the parent given the size of the children and the size of the current vertex.

2. (matchings and induction) Later in the course, we will study matching algorithms that are used in practice to match kidney donors to patients. The challenge in general is that some donors are incompatible with some patients (i.e. the patient's body is likely to reject the donated kidney). Suppose we are very lucky and have n donors and n patients where each donor d is incompatible with exactly one patient, denoted $incomp(d)$, and each patient p is incompatible with exactly one donor $incomp(p)$. (Incompatibility is symmetric, so $incomp(d) = p$ iff $incomp(p) = d$.) Let $f(n)$ be the number of ways, under these circumstances, of matching donors to patients so that each donor donates exactly one kidney to a compatible patient and each patient receives exactly one kidney from a compatible donor.

(a) Show that for all $n \geq 3$,
$$f(n) > (n-1) \cdot f(n-2).$$

Hint: let d be one of the donors, and consider all possible patients p with whom d could be matched. Then consider the case where $incomp(p)$ is matched with $incomp(d)$.

Without loss of generality, consider donors d_1 through d_n and patients p_1 through p_n, where $incomp(d_i) = p_i$ and $incomp(p_i) = d_i$. By this definition, d_1 cannot select p_1, so he has $n-1$ options of patients to match with $p_2 ... p_n$. Conditioning on this, we can consider d_2's options. If d_2 matches with p_1, this leaves $n-2$ remaining donors and patients, allowing us to rerun the original experiment, but updating our input as $n-2$ as opposed to n, which is represented as $f(n-2)$. By multiplication rule, the options we have explored so far are represented by $(n-1) * f(n-2)$. However, since d_2 can also choose to match with a different patient, this will generate even more potential matchings (this case will be formalized in part b), which means that there are even more possibilities than $(n-1) * f(n-2)$, proving that for all $n \geq 3, f(n) > (n-1) * f(n-2)$.

(b) In fact, show that for all $n \geq 3$, we have

$$f(n) = (n-1) \cdot (f(n-1) + f(n-2)).$$

This will require you to include the remaining case where $incomp(p)$ is *not* matched with $incomp(d)$, unlike the previous exercise.
Continuing from part 2a, we start by conditioning on d_1 matching with p_2, leaving him with $(n-1)$ options. From here, we consider the case where d_2 does match with p_1, ensuring that both the d_1, p_2 pair and d_2, p_1 pair are removed from the experiment, resulting in the same experiment being run with $n-2$ donors and patients now (represented as $f(n-2)$). Now, we consider the case where d_2 does not match with p_1 and matches with another patient instead. This will run the same experiment but will exclude the pair d_2, p_1 since they are already matched, but won't assume d_2 is already matched like the previous case did, and this is represented as $f(n-1)$. So if we condition both of these cases on the original matching of the first donor, we get $f(n) = (n-1) * (f(n-1) + f(n-2))$.

(c) Prove by strong induction that for all $n \geq 2$,

$$\frac{n!}{3} \leq f(n) \leq \frac{n!}{2}.$$

We start with our two base cases, $n = 2$ and $n = 3$.

For $n = 2$, we do the following : $\frac{2!}{3} \leq f(2) \leq \frac{2!}{2}$, which simplifies to $\frac{2}{3} \leq f(2) \leq 1$ where $f(2) = 1$ because when there are 2 pairs of donors and patients, there is only one way to match the group, since $incomp(d_i) = incomp(p_i)$, resulting in $\frac{2}{3} \leq 1 \leq 1$, which is true.

For $n = 3$, we can use the equality we proved from part b, which allows us to write : $f(3) = (3-1) \cdot (f(3-1) + f(3-2))$, which becomes $f(3) = (2) \cdot (f(2) + f(1))$, where $f(1) = 0$, since when there is only one patient and one donor, it is assumed that they are incompatible with eachother. We ultimately get $f(3) = (2) \cdot (1+0)$, simplified as $f(3) = 2$

For our inductive hypothesis, we assume $\frac{k!}{3} \leq f(k) \leq \frac{k!}{2}$ where $2 \leq k \leq n$.

For our inductive step, we want to show that $f(n+1) = n * [f(n) + f(n-1)]$. First, we should consider the right side of our inductive hypothesis' inequality $(f(k) \leq \frac{k!}{2})$, when k is substituted with $n+1$, we end up with $f(n+1) \leq n * [\frac{(n)!}{2} + \frac{(n-1)!}{2}]$, which becomes $f(n+1) \leq \frac{n(n)!}{2} + \frac{n(n-1)!}{2}$ which becomes $f(n+1) \leq \frac{n(n)!}{2} + \frac{n!}{2}$ which becomes $f(n+1) \leq \frac{n(n)!+n!}{2}$ which becomes $f(n+1) \leq \frac{n!}{2}(n+1)$ which becomes $f(n+1) \leq \frac{(n+1)!}{2}$.

Next, we should consider the left side of our inductive hypothesis' inequality $(f(k) \geq \frac{k!}{3})$, when k is substituted with $n+1$, we solve this the same we way solved the right side of the inequality, resulting in $f(n+1) \geq \frac{(n+1)!}{3}$. Since both sides of our inequality hold, we know that our inductive step holds. By principle of mathematical induction, we've proved that $\frac{n!}{3} \leq f(n) \leq \frac{n!}{2}$ holds for all $n \geq 2$.

Collaborators : Jada Pisani Lee, Fifi Teklemedhin, Gabriel Korish, Teiga Ajise - Helped unblock me in conceptual understanding of how to think about the problems before solving them.