Lecture VI
AMORTIZATION

Many algorithms amount to a sequence of operations on a data structure. For instance, the well-known **heapsort algorithm** is a sequence of `insert`'s into an initially empty priority queue, followed by a sequence of `deleteMin`'s from the queue until it is empty. Thus if c_i is the cost of the ith operation, the algorithm's running time is $\sum_{i=1}^{2n} c_i$, since there are $2n$ operations for sorting n elements. In worst case analysis, we ensure that *each* operation is efficient, say $c_i = O(\log n)$, leading to the conclusion that the overall algorithm is $O(n \log n)$. The idea of **amortization** exploits the fact that we may be able to obtain the same bound $\sum_{i=1}^{2n} c_i = O(n \log n)$ without ensuring that each c_i is logarithmic. We then say that the **amortized cost** of each operation is logarithmic. Thus "amortized complexity" is a kind of average complexity although it has nothing to do with probability. Tarjan [9] gives the first systematic account of this topic.

Why amortize? Even in problems where we could have ensured *each* operation is logarithmic time, it may be advantageous to achieve only logarithmic behavior in the amortized sense. This is because the extra flexibility of amortized bounds may lead to simpler or more practical algorithms. In fact, many "amortized" data structures are relatively easy to implement. To give a concrete example, consider any balance binary search tree scheme. The algorithms for such trees must perform considerable book-keeping to maintain its balanced shape. In contrast, we will see an amortization scheme for binary search tree which is considerably simpler and "lax" about balancing. The operative word in such amortized data structures is[1] laziness: try to defer the book-keeping work to the future if it can be helped. This will be clearer when we discuss splay trees below.

This lecture is in 3 parts: we begin by introducing the **potential function framework** for doing amortization analysis. Then we introduce two data structures, **splay trees** and **Fibonacci heaps**, which can be analyzed using this framework. We give a non-trivial application of each data structure: splay trees are used to maintain the convex hull of a set of points in the plane, and Fibonacci heaps are used for implement Prim's algorithm for minimum spanning trees.

§1. The Potential Framework

We formulate an approach to amortized analysis using the concept of "potential functions". Borrowing a concept from Physics, we imagine data structures as storing "potential energy" that can be released to do useful work. First, we view a data structure such as a binary search tree as a persistent object that has a state which can be changed by operations (e.g., insert, delete, etc). The *characteristic property* of potential functions is that they are a function of the *current* state of the data structure alone, independent of the history of how the data structure was derived.

A "Counter Example". We begin with a simple example. Suppose that we have a binary counter C that is represented by a linked list of 0's and 1's. The only operation on C is to increment its value. For instance, if $C = (011011)$ then after incrementing, $C = (011100)$. This linked list representation determines our **cost model**: the cost to increment C is defined to be the length of the suffix of C of the form 01^*. (We may assume that C begins with a 0-bit in its binary representation, so a suffix of this form always exists.) Thus in our example, the cost is 3 since C has the suffix 011. The problem we consider is to analyse the cost of a sequence of n increments, starting from an initial counter value of 0. In the worst case, an increment operation costs $\Theta(\lg n)$. Therefore a worst-case analysis would conclude that the total cost is $O(n \lg n)$.

[1] In algorithmics, it appears that we like to turn conventional vices (greediness, laziness, gambling with chance, etc) into virtues.

We can do better by using amortized analysis: let us associate with C a **potential** $\Phi = \Phi(C)$ that is equal to the number of 1's in its list representation. For instance, $\Phi(011011) = 4$. Informally, we will "store" $\Phi(C)$ units of work in C. To analyze the increment operation, we consider two cases. (I) Suppose the least significant bit of C is 0. Then the increment operation just changes this bit to 1. We can **charge** this operation 2 units – one unit to do the work and one unit to pay for the increase in potential. (II) Suppose an increment operation changes a suffix $\underbrace{0111\cdots11}_{k}$ of length $k \geq 2$ into $\underbrace{1000\cdots00}_{k}$: the cost incurred is $\Theta(k)$. Notice that the potential Φ decreases by $k-2$. This decrease "releases" $k-2$ units of work that can pay for $\Theta(k-2)$ of the cost incurred. So we only need to charge this operation 2 units. Thus, in both cases (I) and (II), we only charge 2 units of work for an operation, and so the total charges over n operations is only $2n$. We conclude that the amortized cost of incrementing C is $O(1)$.

Abstract Formulation. We present now one abstract formulation of amortization analysis. It is assumed that we are analyzing the cost of a sequence

$$p_1, p_2, \ldots, p_n$$

of **requests** on a data structure D. The term "request" is meant to cover two types of operations: "updates" that modify D and "queries" that need not[2] modify D. The data structure is dynamically changing: at any moment, the data structure is in some **state**, and each request transforms the current state of D. Let D_i be the state of the data structure after request p_i, with D_0 the initial state.

Each p_i has a non-negative **cost**, denoted $\textsc{Cost}(p_i)$. This cost depends on the complexity model which is part of the problem specification. To carry out an amortization argument, we must specify a **charging scheme** and a **potential function**. Unlike the cost function, the charging scheme and potential function are not inherent to the complexity model. They are artifacts of our analysis and may require some amount of ingenuity to be formulated.

A **charging scheme** is just any systematic way to associate a non-negative number $\textsc{Charge}(p_i)$ to each operation p_i. Informally, we "levy" a **charge** of $\textsc{Charge}(p_i)$ on the operation. We emphasize that this levy need not have any obvious relationship to the cost of p_i. The **credit** of this operation is defined to be the "excess charge",

$$\textsc{Credit}(p_i) := \textsc{Charge}(p_i) - \textsc{Cost}(p_i). \tag{1}$$

In view of this equation, specifying a charging scheme is equivalent to specifying a credit scheme. The credit of an operation can be a negative number (in which case it is really a "debit").

A **potential function** is a non-negative real function Φ on the set of possible states of D satisfying

$$\Phi(D_0) = 0.$$

We call $\Phi(D_i)$ the **potential** of state D_i. The amortization analysis amounts to verifying the following inequality at every step:

$$\textsc{Credit}(p_i) \geq \Phi(D_i) - \Phi(D_{i-1}). \tag{2}$$

We call this the **credit-potential invariant**. We denote the *increase in potential* by

$$\Delta\Phi_i := \Phi(D_i) - \Phi(D_{i-1}).$$

Thus equation (2) can be written: $\textsc{Credit}(p_i) \geq \Delta\Phi_i$.

[2] Nevertheless, it may turn out to be advantageous to modify D into some other data structure (equivalent to D, of course). Thus the state of D can change even in case of a query.

The idea is that credit is stored as "potential" in the data structure.[3] Since the potential function and the charging scheme are defined independently of each other, the truth of the invariant (2) is not a foregone conclusion. It must be verified for each case.

If the credit-potential invariant is verified, we can call the charge for an operation its **amortized cost**. This is justified by the following derivation:

$$\begin{array}{rcll} \sum_{i=1}^n \textsc{Cost}(p_i) & = & \sum_{i=1}^n (\textsc{Charge}(p_i) - \textsc{Credit}(p_i)) & \text{(by the definition of credit)} \\ & \leq & \sum_{i=1}^n \textsc{Charge}(p_i) - \sum_{i=1}^n \Delta\Phi_i & \text{(by the credit-potential invariant)} \\ & = & \sum_{i=1}^n \textsc{Charge}(p_i) - (\Phi(D_n) - \Phi(D_0)) & \text{(telescopy)} \\ & \leq & \sum_{i=1}^n \textsc{Charge}(p_i) & \text{(since } \Phi(D_n) - \Phi(D_0) \geq 0). \end{array}$$

When invariant (2) is a strict inequality, it means that some credit is discarded and the analysis is not tight in this case. For our "counter" example, the invariant is tight in every case! This means that our preceding derivation is an equality at each step until the very last step (when we assume $\Phi(D_n) - \Phi(D_0) \geq 0$). Thus we have the exact cost of incrementing a counter from 0 to n is **exactly** equal to

$$\sum_{i=1}^{n} c_i = 2n - \Phi(D_n)$$

where $\Phi(D_n)$ is the number of 1's in the binary representation of n.

The distinction between "charge" and "amortized cost" should be clearly understood: the former is a definition and the latter is an assertion. A charge can only be called an amortized cost if the overall scheme satisfies the credit-potential invariant.

So what is Amortization? Reviewing the amortization framework, we are given a sequence of n requests on a data structure. We are also given a cost model (this may be implicit) which tells us the true cost c_i for the ith operation. We want to upper bound the total cost $\sum_{i=1}^n c_i$. In the amortization analysis, we hope to achieve a bound that is tighter than what can be achieved by replacing each c_i by the worst case cost. This requires the ability to take advantage of the fact that the cost of each type of request is variable, and depends on the current state of the data structure.

The potential method is a formalized form of amortization analysis where we invent a charging scheme and a potential function. The charging scheme tells us that to charge $\widetilde{c}_i$ for the ith request. After verifying that the credit-potential invariant holds for each operation, we may conclude that the charge is an amortized cost. But the potential function can be generalized in several ways: it need not be defined just for the data structure, but could be defined for any suitable abstract feature. Moreover, the charge for an operation could be split up in several ways, and applied to several potential functions Φ_j $(j = 1, 2, \ldots, k)$. We also do not need to assume that $\Phi_j \geq 0$. If finally $\Phi_j < 0$ we just have to ensure that $-\Phi_j$ to our charges and make sure that

$$\sum_{i=1} c_i \leq \sum_{i=1} \widetilde{c}_i - \sum_{j=1}^{k} \Phi_j.$$

Exercises

[3] Admittedly, we are mixing financial and physical metaphors. The credit or debit ought to be put into a "bank account" and so Φ could be called the "current balance".

Exercise 1.1: We generalize the example of incrementing binary counters. Suppose we have a collection of binary counters, all initialized to 0. We want to perform a sequence of operations, each of the type

$$\texttt{inc}(C), \quad \texttt{double}(C), \quad \texttt{add}(C, C')$$

where C, C' are names of counters. The operation $\texttt{inc}(C)$ increments the counter C by 1; $\texttt{double}(C)$ doubles the counter C; finally, $\texttt{add}(C, C')$ adds the contents of C' to C while simultaneously set the counter C' to zero. Show that this problem has amortized constant cost per operation.

To be precise, we need to define the cost model. The length of a counter is the number of bits used to store its current value (so the length can change). The cost to double a counter C is just 1 (you only need to prepend a single bit to C). The cost of $\texttt{add}(C, C')$ is the number of bits that the standard algorithm needs to look at (and possibly change) when when adding C and C'. E.g., if $C = 11,1001,1101$ and $C' = 110$, then $C + C' = 11,1010,0011$ and the cost is 9. This is because the algorithm only has to look at 6 bits of C and 3 bits of C'. Note that the first 4 bits of C is not looked at (you can think of them being simply "copied" to the output, although this happens by just not doing anything). After this operation, C has the value $11,1010,0011$ and C' has the value 0.
HINT: The potential of a counter C should take into account the number of 1's as well as the bit-length of the counter.

REMARK: in our cost model, $\texttt{add}(C, C')$ *and* $\texttt{add}(C', C)$ *have the same cost.* How to implement this so that our cost model is realistic is left for the next exercise.

♢

Exercise 1.2: In the previous counter problem, we define a cost model for $\texttt{add}(C, C')$ that depends only on the bit patterns in C and C'. In particular, the cost of $\texttt{add}(C, C')$ and $\texttt{add}(C', C)$ are the same. How can you implement this algorithm so that the cost model is realistic?

HINT: To understand the issues, suppose $C = 11,1010,0011$ and $C' = 11$ as in the previous problem. Instead of $\texttt{add}(C, C')$, suppose we want to implement $\texttt{add}(C', C)$. How can you implement this so that the cost of 9 is still realistic? If you simply "add C to C'" in the obvious way, the real cost would be the sum of the lengths of C and C', namely $2 + 10 = 12$. One possibility is to first "add C' to C, then rename these counters". But to implement it this way, you need detect which counter is longer, and to always add the shorter counter to the longer. Another way is to copy the initial results of the addition to an intermediate counter before committing yourself as to which counter will be zero'd out.

♢

Exercise 1.3: Joe Smart says: it stands to reason that if we can increment counters for an amortized cost of $O(1)$, we should be able to also support the operation of "decrementing a counter" in addition to those in the previous exercise. Someone pointed out that the potential functions that have been used so far does not bear out this conjecture of Smart. Smart retorts: of course, the failure of any particular potential function is no proof that my suggestion is incorrect.
(a) Can you please give Joe Smart a more convincing argument?
(b) In what way is the intuition of Joe Smart about the symmetry of decrement and increments correct? Formalize this by a result about amortized cost. ♢

Exercise 1.4: Generalize the previous exercise by assuming that the counters need not be initially zero, but may contain powers of 2. ♢

End Exercises

§2. Splay Trees

The **splay tree data structure** of Sleator and Tarjan [8] is a practical approach to implementing all operations listed in §III.2. A key motivation for splay trees is a simple heuristic called the **move-to-front heuristic** – it basically says that if we want to repeatedly access items in a list, then it is a good idea to move any accessed item to the front of the list, to facilitate future accesses to this item. Of course, there is no guarantee that we would want to access this item again in the future. But even if we never again access this item, we have not lost much because the cost of moving the item has already been paid for (using the appropriate accounting method). Amortization (and probabilistic) analysis can be used to prove that this heuristic is a good idea. This material is in appendix A.

The analogue of the move-to-front heuristic for maintaining binary search trees should be clear: after we access (`lookUp`) a key K in a tree T, we must move it to the root. What if K is not in T? Recall that the **successor** of K in T is the smallest key K' in T such that $K \leq K'$; the **predecessor** of K in T is the largest key K' in T such that $K' \leq K$. Thus K does not have a successor (resp., predecessor) in T if K is larger (resp., smaller) than any key in T. Also, the successor and predecessor coincide with K iff K is in T. We characterize the **splay** operation as follows:

$$\texttt{splay}(\textit{Key } K, \textit{Tree } T) \rightarrow T' \tag{3}$$

re-structures the binary search tree T into an equivalent binary search tree T' so that the key K' at the root of T' is equal to *either* the successor *or* predecessor of K in T. We are indifferent as to whether K' is the successor or predecessor. In particular, if K is smaller than any key in T, then K' is the smallest key in T. A similar remark applies if K is larger than any key in T.

Whenever we use the operation (3) above, the following assumption on T will hold:

T is non-empty and all the keys in T are distinct.

Since T is non-empty, any key K will have a successor or predecessor (perhaps not both) in T and $\texttt{splay}(K, T)$ can always be well-defined. See figure 1 for examples of splaying.

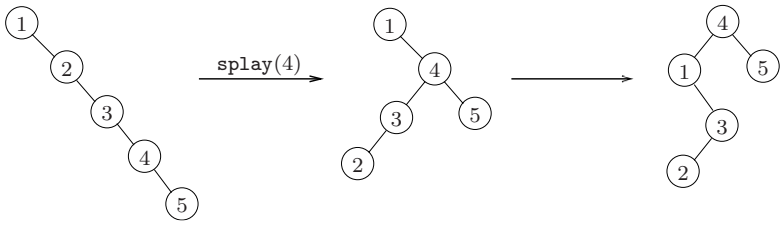

Figure 1: Splaying key 4 (an intermediate step shown).

Before describing the splay algorithm, we show how it will be used.

Reduction to Splaying. We now implement the fully mergeable dictionary ADT (§III.2). The implementation is quite simple: every ADT operation is reducible to one or two splaying operations.

- `lookUp`(*Key* K, *Tree* T): first perform $\texttt{splay}(K, T)$. Then examine the root of the resulting tree to see if K is at the root. `lookUp` is a success iff K is at the root. It is important to realize that we deliberately modify the tree T by splaying it. This is necessary for our analysis.

- `insert`(*Item* X, *Tree* T): again, first perform `splay`$(X.\texttt{Key}, T)$. Then examine the key K' at root of the resulting tree: if $K' = X.\texttt{Key}$, we declare an error (recall that keys are distinct in T). If $K' > X.\texttt{Key}$, we can install a new root containing X, and K' becomes the right child of X as in figure 2. The case $K' < X.\texttt{Key}$ is symmetrical. In either case, the new root has key equal to $X.\texttt{Key}$.
- `merge`(*Tree* T_1, T_2) $\to T$: recall that all the keys in T_1 must be less than any key in T_2. First let $T \leftarrow$ `splay`$(+\infty, T_1)$. Here $+\infty$ denotes an artificial key larger than any real key in T_1. So the root of T has no right child. We then make T_2 the right subtree of T.
- `delete`(*Key* K, *Tree* T): first perform `splay`(K, T). If the root of the resulting tree does not contain K, there is nothing to delete. Otherwise, delete the root and merge the left and right subtrees, as described in the previous bullet.
- `deleteMin`(*Tree* T): we perform $T' \leftarrow$ `splay`$(-\infty, T)$ and return the right subtree of T'.
- `split`(*Key* K, *Tree* T) $\to T'$: perform `splay`(K, T) so that the root of T now contains the successor or predecessor of K in T. Split off the right subtree of T, perhaps including the root of T, into a new tree T'.

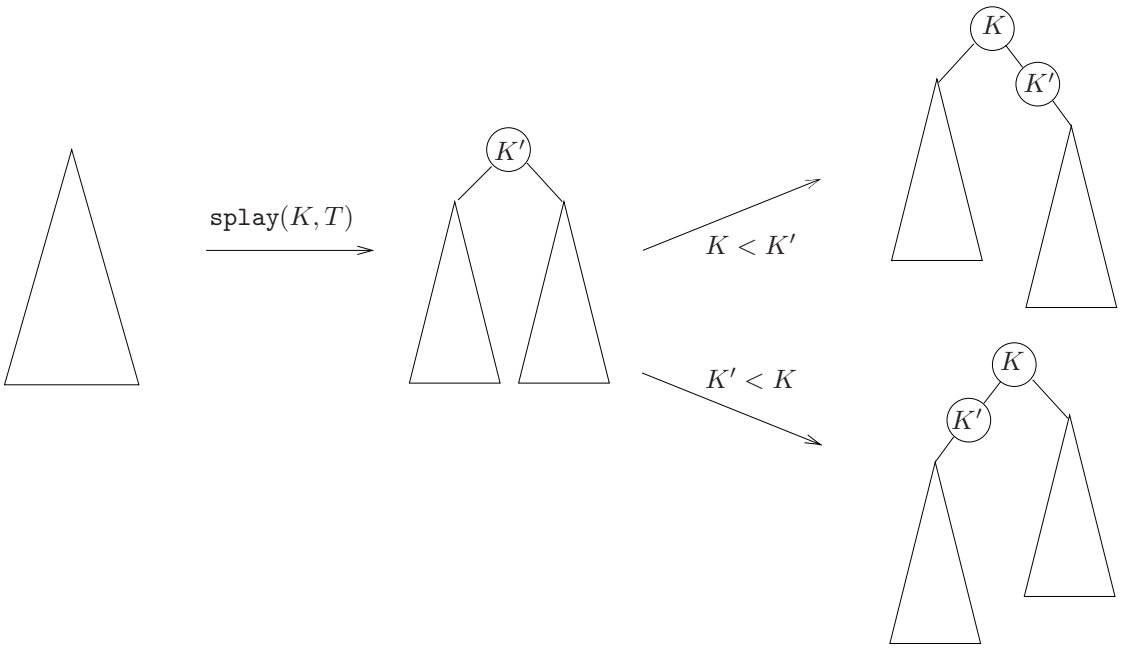

Figure 2: Inserting an item with key K: case $K' > K$.

Reduction to SplayStep. Splaying T at key K is easily accomplished in two stages:

- Perform the usual binary tree search for K. Say we terminate at a node u that contains K in case T contains such a node. Otherwise, let u be the last node that we visit before the binary tree search algorithm attempts to follow a null pointer. This node u contains the successor or predecessor of K in T.
- Now repeatedly call the subroutine

  `splayStep`(u)

  until u becomes the root of T. Termination is guaranteed because `splayStep`(u) always reduce the depth of u.

It remains to explain the SplayStep subroutine. We need a terminology: A grandchild u of a node v is called a **outer left grandchild** if u is the left child of the left child of v. Similarly for **outer right grandchild**. So an **outer grandchild** is either an outer left or outer right grandchild. If a node has a grandparent and is not an outer grandchild, then it is a **inner grandchild**.

`splayStep`(*Node* u):
There are three cases.
Base Case. If u.`Parent` is the root,
then we simply `rotate`(u) (see figure 6).
Case I. Else, if u is an outer grandchild,
perform two rotations: `rotate`(u.`Parent`), followed by `rotate`(u). See figure 3.
Case II. Else, u is an inner grandchild and we perform a double rotation (`rotate`(u) twice). See figure 3.

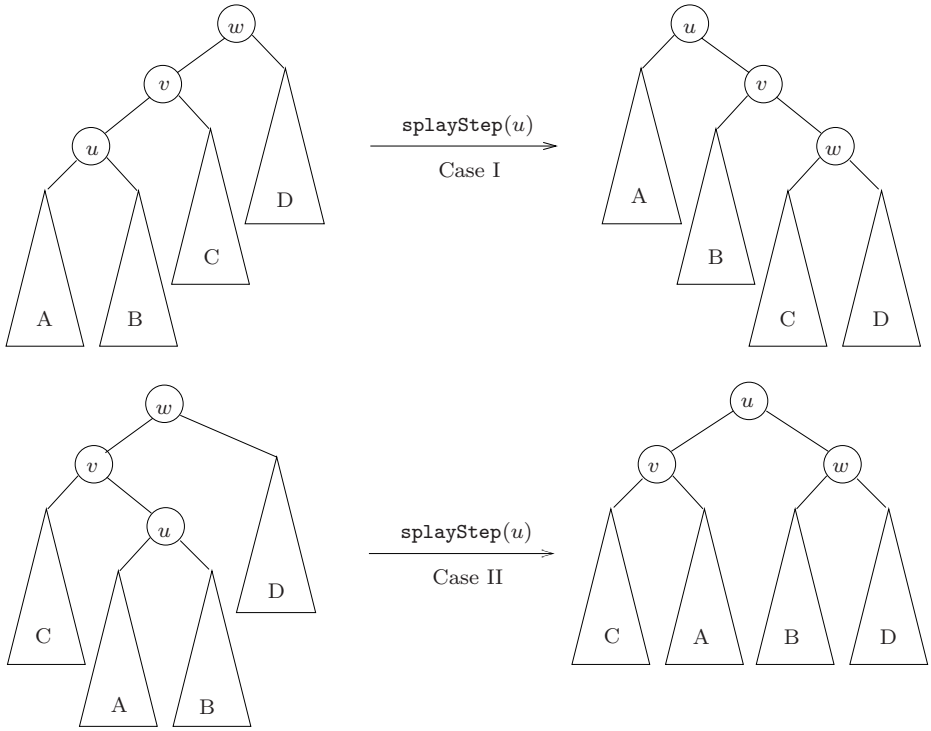

Figure 3: SplayStep at u: Cases I and II.

In figure 1, we see two applications of `splayStep`(4). Sleator and Tarjan calls the three cases of SplayStep the zig (base case), zig-zig (case I) and zig-zag (case II) cases. It is easy to see that the depth of u decreases by 1 in a zig, and decreases by 2 otherwise. Hence, if the depth of u is h, the splay operation will halt in about $h/2$ `splayStep`'s. Recall in §III.6, we call the zig-zag a "double rotation".

We illustrate the fact that `splay`(K, T) may return the successor or predecessor: let T_0 be the splay tree in figure 4. If we call `splay`$(6, T_0)$, the result will be T_1 in the figure, where $u.Key = 7$. But if we call `splay`$(6, T_1)$, the result will be the tree T_2 in the figure, where $u.Key = 5$. What if you call `splay`$(6, T_2)$?

Before moving to the analysis of splay trees, consider the possible behavior of this data structure. Notice that the search trees are by no means required to be balanced. Imagine a sequence of insertions to an empty

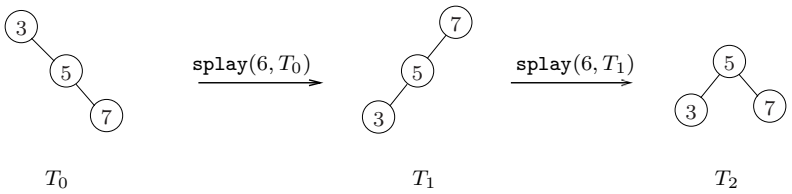

Figure 4: Splaying may return successor or predecessor

tree: if the key of each successive insertion is larger than the previous one, we would end up with a linear structure.

Top Down Splaying. We now introduce a variation of splaying. The Sleator-Tarjan splay algorithms requires two passes over the splay path. Suppose we wish to have a one-pass algorithm. The basic idea is this: *for each node that we "visit" in our search path, we will make it the root before we "visit" it.* Initially, we begin at the root so the basic idea is satisfied. The next node (if any) we visit is either the left or right child of the root. So our basic idea amounts to just performing a left or right rotation at the root, and now we continue recursively. This idea has a pitfall (Exercise). The correct solution is as follows. Let the top-down splaying procedure be denoted `topSplay`($KeyK, Nodeu$). We have 4 possible states of our algorithm:

- State 0: Both u_L and u_R have not been visited.
- State 1: u_L but not u_R has been visited.
- State 2: u_R but not u_L has been visited.
- State 3: Both u_L and u_R have been visited.

Here is the transition rule for the states.

State 0: Initially, we are in state 0. If u.`Key` $> K$, then rotate u.`Left` and we move into state 1; if u.`Key` $< K$, then we rotate u.`Right` and we move into state 2.

State 1: If u.`Key` $> K$ then we next move into state 3 and perform the actions

$$v \leftarrow u\texttt{.Left.Right}; \mathit{rotate}(v); \mathit{rotate}(v); \texttt{topSplay}(K, v).$$

Otherwise, u.`Key` $< K$ and remain in state 1 and perform the actions

$$v \leftarrow u\texttt{.Right}; \texttt{rotate}(v); \texttt{topSplay}(K, v).$$

State 2: This is symmetrical to State 1.

State 3: Once we are in state 3, we remain in state 3. If u.`Key` $> K$ then $v \leftarrow u$.`Left.Right` else $v \leftarrow u$.`Right.Left`. In any case, perform the actions

$$\mathit{rotate}(v); \mathit{rotate}(v); \texttt{topSplay}(K, v).$$

An alternative description is to perform cases I, II or III in direct analogy to SplayStep.

Exercises

Exercise 2.1: Perform the following splay tree operations, starting from an initially empty tree.

$$Ins(3,2,1,6,5,4,9,8,7), LookUp(3), Del(7), Ins(12,15,14,13), Split(8).$$

Show the splay tree after each step. ♢

Exercise 2.2: Show the result of `merge`(T_1, T_2) where T_1, T_2 are the splay trees shown in figure 5. ♢

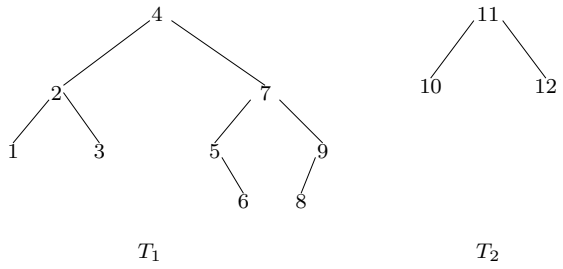

Figure 5: Splay trees T_1, T_2

Exercise 2.3: Show that the worst case time for any of the splay tree operations is $\Omega(n)$. ♢

Exercise 2.4: To splay a tree at a key K, our algorithm begins by doing the conventional `lookUp` of K. If K is not in the tree, and u is the last node reached, then clearly u has at most one child. Prove that u contains the successor or predecessor of K. ♢

Exercise 2.5: Let T be a binary search tree in which every non-leaf has one child. Thus T has a linear structure with a unique leaf.
(a) What is the effect of `lookUp` on the key at the leaf of T?
(b) What is the minimum number of `lookUp`'s to make T balanced? ♢

Exercise 2.6: A variant of the insertion algorithm is to make the inserted node to be equal to either the left or right child of the root. What are the relative advantages/disadvantages of this over what we specified in the text? ♢

Exercise 2.7: (Top Down Splaying)
(a) Explain the "pitfall" mentioned for the obvious implementation of the top-down splaying algorithm.
(b) Give an efficient implementation of `topSplay` as described above. Efficiency here means trying to reduce the number of pointer manipulations, and this may entail combining the pointer manipulations of several rotations.
(c) Do an empirical study of Top Down Splaying, comparing its performance to standard Splaying. ♢

Exercise 2.8: A **splay tree** is a binary search tree that arises from a sequence of splay tree operations, starting from empty trees.
(a) Is every binary search tree a splay tree?
(b) Let T and T' be equivalent binary search trees (i.e., they store the same set of keys). Can we transform T to T' by repeated splays? ♢

End Exercises

§3. Splay Analysis

Our main goal next is to prove:

(*) *The amortized cost of each splay operation is* $O(\log n)$ *assuming at most* n *items in a tree.*

Let Size(u) denote as usual the number of nodes in the subtree rooted at u, and define its **potential** to be

$$\Phi(u) = \lfloor \lg \textsc{Size}(u) \rfloor .$$

Initially, the data structure has no items and has zero potential. If $S = \{u_1, u_2, \ldots\}$ is a set of nodes, we may write $\Phi(S)$ or $\Phi(u_1, u_2, \ldots)$ for the sum $\sum_{u \in S} \Phi(u)$. If S is the set of nodes in a splay tree T or in the entire data structure then $\Phi(S)$ is called the potential of (respectively) T or the entire data structure.

Lemma 1 (key) *Let* Φ *be the potential function before we apply* `splayStep`*(*u*), and let* Φ' *be the potential after. The credit-potential invariant is preserved if we charge the SplayStep*

$$3(\Phi'(u) - \Phi(u)) \tag{4}$$

units of work in cases I and II. In the base case, we charge one extra unit, in addition to the charge (4).

The main goal (*) follows easily from this key lemma. To see this, suppose that splaying at u reduces to a sequence of k SplaySteps at u and let $\Phi_i(u)$ be the potential of u after the ith SplayStep. The total charges to this sequence of SplaySteps is

$$1 + \sum_{i=1}^{k} 3[\Phi_i(u) - \Phi_{i-1}(u)] = 1 + 3[\Phi_k(u) - \Phi_0(u)]$$

by telescopy. Note that the "1" comes from the fact that the last SplayStep may belong to the base case. Clearly this total charge is at most $1 + 3 \lg n$. To finish off the argument, we must account for the cost of looking up u. But it easy to see that this cost is proportional to k and so it can be covered by charging one extra unit to every SplayStep. This only affects the constant factor in our charging scheme. This concludes the proof of the main goal.

We now address address the Key Lemma. The following is a useful remark about rotations:

Lemma 2 *Let* Φ *be the potential function before a rotation at* u *and* Φ' *the potential function after. Then the increase in potential of the overall data structure is at most*

$$\Phi'(u) - \Phi(u).$$

The expression $\Phi'(u) - \Phi(u)$ *is always non-negative.*

Proof. We refer to figure 6. The increase in potential is

$$\begin{array}{rcll} \Delta\Phi & = & \Phi'(u,v) - \Phi(u,v) & \\ & = & \Phi'(v) - \Phi(u) & (\text{as } \Phi'(u) = \Phi(v)) \\ & \leq & \Phi'(u) - \Phi(u) & (\text{as } \Phi'(u) \geq \Phi'(v)). \end{array}$$

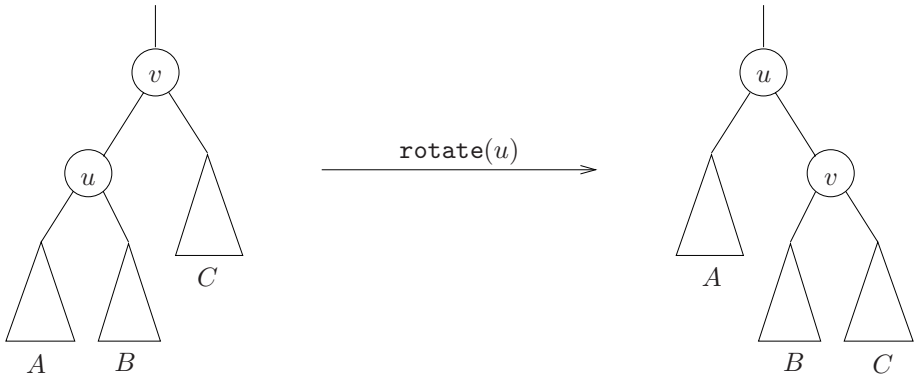

Figure 6: Rotation at u.

It is obvious that $\Phi'(u) \geq \Phi(u)$. **Q.E.D.**

Proof of Key Lemma. The base case is almost immediate from lemma 2: the increase in potential is at most $\Phi'(u) - \Phi(u)$. This is at most $3(\Phi'(u) - \Phi(u))$ since $\Phi'(u) - \Phi(u)$ is non-negative. The charge of $1 + 3(\Phi'(u) - \Phi(u))$ can therefore pay for the cost of this rotation and any increase in potential.

Refer to figure 3 for the remaining two cases. Let the sizes of the subtrees A, B, C, D be a, b, c, d, respectively.

Consider case I. The increase in potential is

$$\begin{array}{rcll} \Delta\Phi & = & \Phi'(u,v,w) - \Phi(u,v,w) & \\ & = & \Phi'(v,w) - \Phi(u,v) & (\text{as } \Phi'(u) = \Phi(w)) \\ & \leq & 2(\Phi'(u) - \Phi(u)) & (\text{as } 2\Phi'(u) \geq \Phi'(v,w), \quad 2\Phi(u) \leq \Phi(u,v)). \end{array}$$

Since $\Phi'(u) \geq \Phi(u)$, we have two possibilities: (a) If $\Phi'(u) > \Phi(u)$, then the charge of $3(\Phi'(u) - \Phi(u))$ can pay for the increased potential *and* the cost of this splay step. (b) Next suppose $\Phi'(u) = \Phi(u)$. By assumption, $\Phi'(u) = \lfloor \lg(3 + a + b + c + d) \rfloor$ and $\Phi(u) = \lfloor \lg(1 + a + b) \rfloor$ are equal. Thus $1 + a + b > 2 + c + d$, and so $3 + a + b + c + d > 2(2 + c + d)$ and

$$\Phi'(w) = \lfloor \lg(1 + c + d) \rfloor < \lfloor \lg(3 + a + b + c + d) \rfloor = \Phi(u).$$

Also,

$$\Phi'(v) \leq \Phi'(u) = \Phi(u) \leq \Phi(v).$$

Combining these two inequalities, we conclude that

$$\Phi'(w, v) < \Phi(u, v).$$

Hence $\Delta\Phi = \Phi'(w, v) - \Phi(u, v) < 0$. Since potentials are integer-valued, this means that $\Delta\Phi \leq -1$. Thus the change in potential releases at least one unit of work to pay for the cost of the splay step. Note that in this case, we charge nothing since $3(\Phi'(u) - \Phi(u)) = 0$. Thus the credit-potential invariant holds.

Consider case II. The increase in potential is again $\Delta\Phi = \Phi'(v, w) - \Phi(u, v)$. Since $\Phi'(v) \leq \Phi(v)$ and $\Phi'(w) \leq \Phi'(u)$, we get

$$\Delta\Phi \leq \Phi'(u) - \Phi(u).$$

If $\Phi'(u) - \Phi(u) > 0$, then our charge of $3(\Phi'(u) - \Phi(u))$ can pay for the increase in potential and the cost of this splay step. Hence we may assume otherwise and let $t = \Phi'(u) = \Phi(u)$. In this case, our charge is

$3(\Phi'(u) - \Phi(u)) = 0$, and for the credit potential invariant to hold, it suffices to show

$$\Delta\Phi < 0.$$

It is easy to see that $\Phi(v) = t$, and so $\Phi(u, v) = 2t$. Clearly, $\Phi'(v, w) \leq 2\Phi'(u) = 2t$. If $\Phi'(v, w) < 2t$, then $\Delta\Phi = \Phi'(v, w) - \Phi(u, v) < 0$ as desired. So it remains to show that $\Phi'(v, w) = 2t$ is impossible. For, if $\Phi'(v, w) = 2t$ then $\Phi'(v) = \Phi'(w) = t$ (since $\Phi'(v), \Phi'(w)$ are both no larger than t). But then

$$\Phi'(u) = \lfloor \lg(\textsc{Size}'(v) + \textsc{Size}'(w) + 1) \rfloor \geq \lfloor \lg(2^t + 2^t + 1) \rfloor \geq t + 1,$$

a contradiction. This proves the Key Lemma.

We conclude with the main result on splay trees.

Theorem 3 *A sequence of m splay tree requests (*`lookUp`*,* `insert`*,* `merge`*,* `delete`*,* `split`*) involving a total of n items takes $O(m \log n)$ time to process. As usual, we assume that the potential of the data structure is initially* 0*.*

Proof. This follows almost immediately from (*) since each request can be reduced to a constant number of splay operations plus $O(1)$ extra work. We need to attend to one detail in Insertion. Here, we introduce a new node with potential at most $\lg n$. This increase of potential must be charged but clearly this additional does not change our overall cost. Similarly for Merge and Deletion. **Q.E.D.**

Sherk [7] has generalized splaying to k-ary search trees. In such trees, each node stores an ordered sequence of $t - 1$ keys and t pointers to children where $2 \leq t \leq k$. This is similar to B-trees.

Application: Splaysort Clearly we can obtain a sorting algorithm by repeated insertion into a splay tree. Such an algorithm has been implemented [5]. Splaysort has the ability to take advantage of "presortedness" in the input sequence and hence may run faster than Quicksort for some inputs. One way to quantify presortedness is to count the number of pairwise inversions in the input sequence.

Exercises

Exercise 3.1: Where in the proof is the constant "3" actually needed in our charge of $3(\Phi'(u) - \Phi(u))$? ♢

Exercise 3.2: Adapt the proof of the Key Lemma to justify the following variation of SplayStep:

> VarSplayStep(u):
> (Base Case) if u is a child or grandchild of the root,
> then rotate once or twice at u until it becomes the root.
> (General Case) else rotate at u.`Parent`, followed by two rotations at u.

♢

Exercise 3.3:
(i) Is it true that splays always decrease the height of a tree? The average height of a tree? (Define

the average height to be the average depth of the leaves.)
(ii) What is the effect of splay on the last node of a binary tree that has a linear structure, *i.e.*, in which every internal node has only one child? HINT: First consider two simple cases, where all non-roots is a left child and where each non-root is alternately a left child and a right child. ♢

Exercise 3.4: Assume that node u has a great-grandparent. Give a simple description of the effect of the following sequence of three rotations: `rotate`(u.`Parent.Parent`); `rotate`(u.`Parent`); `rotate`(u). ♢

Exercise 3.5: For any node u,

$$\Phi(u_L) = \Phi(u_R) \Rightarrow \Phi(u) = \Phi(u_L) + 1$$

where u_L, u_R are the left and right child of u. ♢

Exercise 3.6: Modify our splay trees to maintain (in addition to the usual children and parent pointers) pointers to the successor and predecessor of each node. Show that this can be done without affecting the asympotic complexity of all the operations (`lookUp`, `insert`, `delete`, `merge`, `split`) of splay trees. ♢

Exercise 3.7: We consider some possible simplifications of the `splayStep`.
(A) One-rotation version: Let `splayStep`(u) simply amount to `rotate`(u).
(B) Two-rotation version:

> SplayStep(u):
> (Base Case) if u.`Parent` is the root, `rotate`(u).
> (General Case) else do `rotate`(u.`Parent`), followed by `rotate`(u).

For both (A) and (B):
(i) Indicate how the proposed SplayStep algorithm differs from the original.
(ii) Give a general counter example showing that this variation does not permit a result similar to the Key Lemma. ♢

Exercise 3.8: Modify the above algorithms so that we allow the search trees to have identical keys. Make reasonable conventions about semantics, such as what it means to lookup a key. ♢

Exercise 3.9: Can we use the simpler potential function $\Phi(u) = \lg \text{Size}(u)$ in our splay analysis? ♢

End Exercises

§4. Application to Convex Hulls

The following application is interesting because it illustrates the idea of an **implicit binary search tree**. The usual notion of keys is inapplicable. But by using information distributed at a node u and its children u_L and u_R, we are able to perform tests to make decision which simulates searching in a binary search tree.

Given a finite set X of points in the plane, its **convex hull** $CH(X)$ is the smallest convex subset of the plane that contains X. As $CH(X)$ is a convex polygon, we may represent it as a sequence

$$H = (v_1, v_2, \ldots, v_n)$$

where $v_i \in X$ and these v_i's appear as consecutive vertices of the polygon $CH(X)$. We shall use H and $CH(X)$ interchangeably. We want to dynamically maintain H subject to two types of requests:

$$\texttt{tangent}(p, H) \quad \text{and} \quad \texttt{insert}(p, H)$$

where p is a new point. If p is outside H, $\texttt{tangent}(p, H)$ will return a pair (q, r) of distinct points on H such that the lines $\overline{pq}$ and $\overline{pr}$ are both tangential to H. E.g., in figure 7(a), v_3, v_5 are the tangent points from p.

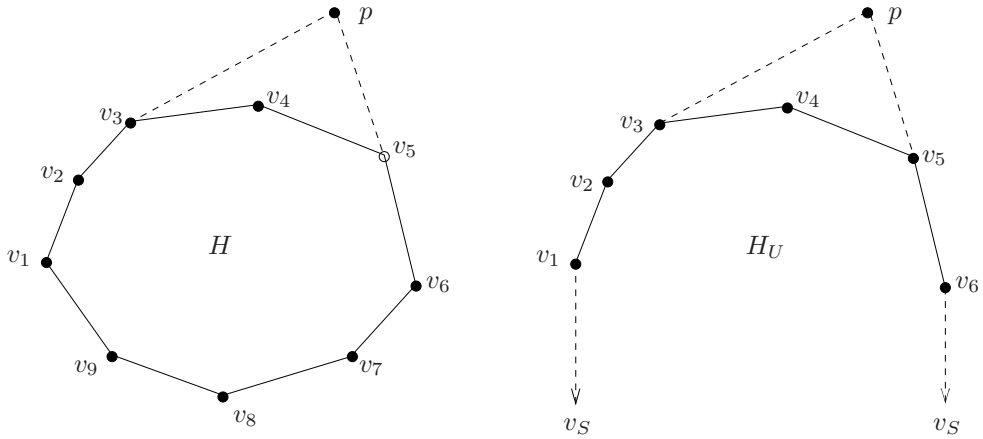

Figure 7: (a) $H = (v_1, \ldots, v_9)$, (b) $H_U = (v_1, \ldots, v_6)$.

We call q and r the **tangent points** of H from p. If p is inside the current hull, we return "↑" since H has no tangent points from p. The request $\texttt{insert}(p, H)$ means we want to update H to represent $CH(X \cup \{p\})$. Note that if p is inside the current hull, H is unchanged.

Reduction to Half-Hulls. We may assume that v_1 and v_ℓ $(1 \le \ell \le n)$ has the smallest and largest x-coordinates among the v_i's (ties are arbitrarily broken). We can break H into two convex chains,

$$H_U = (v_1, v_2, \ldots, v_\ell), \qquad H_L = (v_1, v_n, v_{n-1}, \ldots, v_{\ell+1}, v_\ell).$$

Note that H_U and H_L share precisely their common endpoints. Assuming that H_U lies above the segment $v_1 v_\ell$, we call H_U the **upper chain** and H_L the **lower chain** of H. Let $v_S = (0, -\infty)$ and $v_N = (0, +\infty)$ be points[4] at infinity (the 'south' and 'north poles', respectively). We can also think of H_U as the convex hull of $X \cup \{v_S\}$, which strictly speaking should be represented as $(v_1, v_2, \ldots, v_\ell, v_S)$. In general, the convex hull of $X \cup \{v_S\}$ is called the **upper hull** of X. Similarly, we also think of H_L as the convex hull of $X \cup \{v_N\}$ or **lower hull** of X. Collectively, we may call H_U and H_L **half-hulls**.

By symmetry, we henceforth focus on upper hulls. The requests $\texttt{tangent}(p, H_U)$ and $\texttt{insert}(p, H_U)$ can be defined in the natural way. We can implement $\texttt{insert}(p, H)$ simply as

$$\texttt{insert}(p, H_U); \texttt{insert}(p, H_L).$$

[4]Actually, v_S and v_N can be defined to be $(r, -\infty)$ and $(r', +\infty)$ for any finite real numbers r, r'.

Clearly p is inside H iff p is inside both H_U and H_L. Now suppose p is not inside $H_U = (v_1, \ldots, v_\ell)$. Then $\texttt{tangent}(p, H_U)$ returns the pair (q, r) of tangent points of H_U from p where q lies to the left of r. For instance, $\texttt{tangent}(p)$ returns v_3, v_5 in figure 7. There are two special cases. If p is left of v_1, then $q = v_S$; if p is right of v_ℓ then $r = v_S$. The details of how to reduce $\texttt{tangent}(p, H)$ to half-hulls is left to an exercise.

Reduction to Fully Mergeable Dictionary Operations. The fully mergeable dictionary ADT was introduced in §III.2. Let us now assume that the upper hull H_U is stored in a splay tree T using the x-coordinates of vertices as keys. We implement the requests

$$\texttt{insert}(p, T) \text{and } \texttt{tangent}(p, T).$$

We assume that nodes in T have successor and predecessor pointers.

To find the tangent points (q, r) from a query point $p = (p.x, p.y)$, we first do a lookup on $p.x$. Suppose as a result of this lookup, we determine the two consecutive hull vertices v_i, v_{i+1} such that

$$v_i.x < p.x \leq v_{i+1}.x \qquad (1 \leq i < \ell). \tag{5}$$

We can then decide if p is inside the upper hull or not — this amounts to whether p is below the line $\overline{v_i, v_{i+1}}$ or not. If inside, we return $\uparrow$. Otherwise, we want to return the pair (q, r) of tangent points from p. Consider how we locate q (locating r is similar). We know that $q = v_{i_0}$ for some $i_0 \leq i$. Moreover, for any point v_k where $k \leq i$, we can decide whether $k = i_0$, $k < i_0$ or $k > i_0$ according to the following cases:

(i) v_{k-1} **and** v_{k+1} **lie below** $\overline{v_k p}$**.**
Then $i_0 = k$. Here, if $k = 1$ then v_0 is the south pole (which lies below any line).

(ii) v_{k-1}**, but not** v_{k+1}**, lies below** $\overline{v_k p}$**.**
Then $i_0 > k$.

(iii) v_{k+1}**, but not** v_{k-1}**, lies below** $\overline{v_k p}$**.**
Then $i_0 < k$.

We can use this 3-way decision to perform an "implicit binary search" for $q = v_{i_0}$ as follows:

FindLeftTangentPoint(p, T):

1. Initialize u to the root of T.
2. Repeat:
   Let v_k $(1 \leq k < \ell)$ be the vertex stored at u.
   If $v_k.x \geq p.x$, set $u \leftarrow u.\texttt{leftChild}$.
   Else, we have the three possibilities (cases i-iii above):
   (Case i) return(v_k).
   (Case ii) Set $u \leftarrow u.\texttt{rightChild}$.
   (Case iii) Set $u \leftarrow u.\texttt{leftChild}$.

If there is no index i such that (5) holds, then either $p.x \leq v_1.x$ or $p.x > v_\ell.x$. We leave the details for this case to the reader.

Next consider the implementation of $\texttt{insert}(p, T)$. We first perform $\texttt{tangent}(p, T)$ and assume the non-trivial case where a pair (q, r) of tangent points are returned. Then we need to delete from T those vertices

v_i that lies strictly between q and r, and replace them by the point p. This is easily accomplished using the operations of `split` and `merge` on splay trees. This is left for an exercise.

We conclude with the following. Let D be our data structure for the convex hull H (so D is basically a pair of splay trees).

Theorem 4
(i) Using the data structure D to represent the convex hull H of a set of points, we can support `insert`(p, D) *and* `tangent`(p, D) *requests with an amortized cost of $O(\log |H|)$ time per request.*
(ii) From D, we can produce the cyclic order of points in H in time $O(|H|)$. In particular, this gives an $O(n \log n)$ algorithm for computing the convex hull of a set of n points.

Our data structure D for representing convex hulls is only semi-dynamic because we do not support the deletion of points. If we want to allow deletion of points, then points that are inside the current convex hull must be represented in the data structure. Overmars and van Leeuwen designed a data structure for a fully dynamic convex hull that uses $O(\log^2 n)$ time for insertion and deletion.

Exercises

Exercise 4.1: Let $a, b, c \in \mathbb{R}^2$.
(i) Show that if

$$M = \begin{bmatrix} a_x & a_y & 1 \\ b_x & b_y & 1 \\ c_x & c_y & 1 \end{bmatrix}$$

then $\det M$ is twice the signed area of the triangle $\Delta(a, b, c)$. Thus, a, b, c are collinear or coincident iff $\det M = 0$. Also, show that $\det M > 0$ iff a, b, c list the vertices counter-clockwise about the triangle.
(ii) Let R be the smallest axes-parallel triangle that contains $\Delta(a, b, c)$. Then at least one of the vertices of $\Delta(a, b, c)$ must be at a corner of R. Without loss of generality, let a be the south-west corner of R, b touches the right vertical edge of R and c touches the top horizontal edge of R. Let r be the rectangle with one corner at a and whose opposite corner is (c_x, b_y). Show by a direct geometric argument that the area of $\Delta(a, b, c)$ is equal to $(|R| - |r|)/2$ where $|R|, |r|$ are the areas of the rectangles R, r (respectively). Hence $|R| - |r|$ is the area of the "L" shape $R \setminus r$.
(iii) Verify that the result of (ii) also follows from (i). ♢

Exercise 4.2: Prove missing details in the above description.
(i) Remaining cases in implementation of `tangent`(p, H_U).
(ii) Implementation of `tangent`(p, H) in terms of `tangent`(p, H_U) and `tangent`(p, H_L).
(iii) Implementation of `insert`(p, H_U). ♢

Exercise 4.3: Suppose that we do not need to implement the `tangent`$(\cdot, \cdot)$ query. This would be the case if we are only interested in constructing the convex hull. Show that we can achieve the same $O(\log n)$ per `insert`$(\cdot, \cdot)$ by a simpler algorithm. Specifically, we can avoid the implicit binary search procedure. ♢

End Exercises

§5. Fibonacci Heaps

The **Fibonacci heap data structure** invented by Fredman and Tarjan gives an efficient implementation of the mergeable queues abstract data type (ADT), which we now explain.

The mergeable queues ADT. The mergeable queues ADT involves domains of three types: *Key*, *Item* and *(mergeable) Queue*. As usual, each item stores a key and each queue stores a collection of items. The ADT represents a collection of queues, supporting these operations:

`makeQueue`()→ Q	returns an empty queue Q
`insert`(*Item* x, *Queue* Q)	
`union`(*Queue* Q_1, Q_2)	
`deleteMin`(*Queue* Q)→ *Item* x	x is minimum item in Q, which is now deleted
`decreaseKey`(*Item* x, *Key* k, *Queue* Q).	

Mergeable queues are clearly extensions of priority queues (§III.2). The above operations are mostly self-explanatory. In the union of Q_1, Q_2, the items in Q_2 are first moved into queue Q_1, then queue Q_2 is destroyed. Thus, the number of queues can increase or decrease over the lifetime of the data structure. The operation `deleteMin`(Q) returns a minimum item in Q, and this item is deleted from Q. This operation is unspecified in case Q is empty. In `decreaseKey`(x, k, Q), we make k the new key of x in Q. But this operation assumes k is smaller than the current key of x – otherwise, we may define it to be either an error or a null-operation (we will leave this decision unspecified).

There may be useful operations that should be provided in practice but omitted above for the sake of economy: deleting an item, making a singleton queue, getting the minimum item without deleting it. These can be defined as follows:

`delete`(*Item* x, *Queue* Q)	≡	`decreaseKey`(x,$-\infty$,Q); `deleteMin`(Q).
`makeQueue`(*Item* x)→ Q	≡	`makeQueue`()→ Q ; `insert`(x,Q).
`Min`(*Queue* Q)→ x	≡	`deleteMin`(Q)→ x; `insert`(x,Q).

The Fibonacci heap data structure. Each mergeable queue is implemented by a Fibonacci heap. A Fibonacci heap H is a collection of trees $T_1, \ldots, T_m$ with these properties:

- Each tree T_i satisfies the min-heap property. In particular, the root of T_i has the minimum item in T_i.
- The roots of these trees are kept in a doubly-linked list, called the **root-list** of H.
- There are two fields H.`Min`, $H.n$ associated with H. The field H.`Min` points to the node with a minimum key, and $H.n$ is the number of items in H.
- For each node x in a tree T_i, we have four pointers that point to (i) the parent of x, (ii) one of its children, and (iii) two of its siblings. The sibling pointers are arranged so that all the children of x appears in a circular doubly-linked list called the **child-list** of x. If y is a child of x, the **sibling-list** of y is the child-list of x. Also, we keep track of x.`degree` (the number of children of x) and x.`mark` (a Boolean value to be explained).

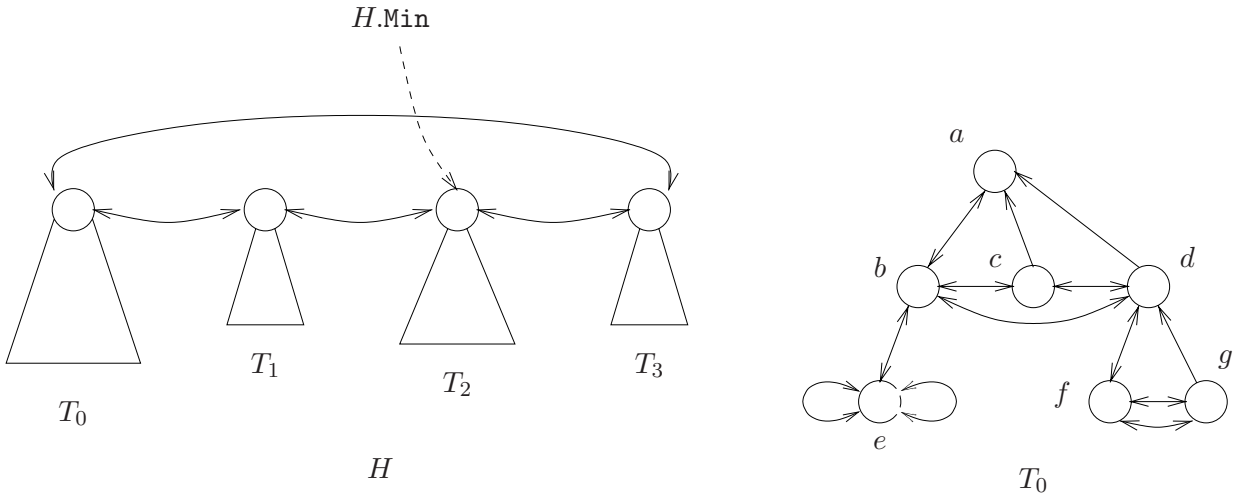

Figure 8: A Fibonacci heap $H = (T_0, \ldots, T_3)$: T_0 in detail

This is illustrated in figure 8. One of the trees T_0 is shown in detail: the root a of T_0 has 3 children b, c and d and they each point to a; on the other hand, a points only to b. There are two non-trivial sibling lists: (b, c, d) and (f, g).

Linking, cutting and marking. We describe some elementary operations used in maintaining Fibonacci heaps.

(a) If x, y are two roots such that the item in x is not less than the item in y then we can **link** x **and** y: this simply makes y the parent of x. The appropriate fields and structures are updated. E.g., x is deleted from the root-list, but inserted into the child-list of y, the degree of y incremented, etc. This operation costs $O(1)$.

(b) The converse to linking is **cutting**. If x is a non-root in a Fibonacci heap H then we can perform *Cut*(x,H): this basically removes x from the child-list of its parent and inserts x into the root-list of H. The appropriate data variables are updated. E.g., the degree of the parent of x is decremented. Again, this operation costs $O(1)$.

(c) We say x is **marked** if $x.\texttt{mark} = true$, and **unmarked** otherwise. Initially, x is unmarked. Our rules will ensure that a root is always unmarked. We mark x if x is not a root and x loses a child (*i.e.*, a child of x is cut); we unmark x when x itself is cut (and put in the root-list). Moreover, we ensure that a marked x does not lose another child before x itself is cut (thereby reverting to unmarked status).

To do amortized analysis, we define a potential function. The **potential** of a Fibonacci heap H is defined as

$$\Phi(H) := t(H) + 2 \cdot m(H)$$

where $t(H)$ is the number of trees in H and $m(H)$ is the number marked items in H. The potential of a collection of Fibonacci heaps is just the sum of the potentials of the individual heaps.

One more definition: let $D(n)$ denote the maximum degree of a node in a Fibonacci heap with n items. We will show later that $D(n) \leq 2 \lg n$.

Remark: The reader may observe how "low-tech" this data structure appears – along with the humble array structure, linked-lists is among the simplest data structures. Yet we intend to achieve the best known overall performance for mergeable queues with Fibonacci heaps. This should be viewed as a testimony to the power of amortization.

§6. Fibonacci Heap Algorithms

We now implement the mergeable queue operations. Our goal is to achieve an amortized cost of $O(1)$ for each operation except for `deleteMin`, which will have logarithmic amortized cost.

Recall that for each operation p, we have a **cost** $\textsc{Cost}(p)$ which will be mostly self-evident in the following description. We must define a **charge** $\textsc{Charge}(p)$. The **credit** is thereby determined: $\textsc{Credit}(p) = \textsc{Charge}(p) - \textsc{Cost}(p)$. This charging scheme will achieve the stated goal in the previous paragraph: $\Theta(1)$ charges for all the non-deletion operations, and $\Theta(\log n)$ for the two deletion operations. Finally, we verify the credit-potential invariant equation (2) for each operation.

`makeQueue`(): we create an empty root-list. The cost is 1, the charge is 1, so credit is 0, and finally $\Delta\Phi = 0$. The credit-potential invariant holds trivially.

> The cost and $\Delta\Phi$ is automatic at this point (our earlier decisions have determined this). Although we said that the "charge" is part of our creative design, at this point, we really have little choice if we wish to satisfy the credit-potential invariant. We might as well define charge to be (at least) the cost plus $\Delta\Phi$.

`insert`(H, x): we create a new tree T containing only x and insert T into the root-list of H. Update H.`Min`, etc. Let us check the credit-potential invariant:

$$\textsc{Cost} \leq 1, \quad \textsc{Charge} = 2, \quad \textsc{Credit} \geq 1, \quad \Delta\Phi = 1.$$

`union`(H_1, H_2): concatenate the two root-lists and call it H_1. Update $\min[H_1]$, etc. Checking the credit-potential invariant:

$$\textsc{Cost} \leq 1, \quad \textsc{Charge} = 1, \quad \textsc{Credit} \geq 0, \quad \Delta\Phi = 0.$$

`deleteMin`(H): we remove H.`Min` from the root-list, and the child-list of H.`Min` can now be regarded as the root-list of another Fibonacci heap. These two circular lists can be concatenated in constant time into a new root-list for H. If t_0 is the old value of $t(H)$, the new value of $t(H)$ is at most $t_0 + D(n)$. Next we need to find the new value of H.`Min`. Unfortunately, we do not know the new minimum item of H. There is no choice but to scan the new root-list of H. While scanning, we might as well[5] spend some extra effort to save future work. This is a process called **consolidation** which is explained next.

Consolidation. In this process, we are given a root-list of length L ($L \leq t_0 + D(n)$ above). We must visit every member in the root-list, and at the same time do repeated linkings *until there is at most one root of each degree*. We want to do this in $O(L)$ time. By assumption, each root has degree at most $D(n)$.

The basic method is that, for each root x, we try to find another root y of the same degree and link the two. So we create a 'new' root of degree $k+1$ from two roots of degree k. If we detect another root of

[5] OK, we may be lazy but not stupid.

degree $k+1$, we link these two to create another 'new' root of degree $k+2$, and so on. The way that we detect the presence of another root of the same degree is by indexing into an array $A[1..D(n)]$ of pointers. Initialize all entries of the array to nil. Then we scan each item x in the root-list. If $k = x.\texttt{degree}$ then we try to "insert" x into $A[k]$. This means making $A[k]$ point to x. But we only do this if $A[x.\texttt{degree}] = \text{nil}$; in case $A[x.\texttt{degree}] \neq \text{nil}$, then it points to some y. In this case, link x to y or vice-versa. If x is linked to y, the latter now has degree $k+1$ and we try to "insert" y into $A[k+1]$, and so on. So each failed insertion leads to a linking, and there are at most L linking operations. Since each linking removes one root, there are at most L linkings in all. (This may not be obvious if we see this the wrong way!) Thus the total cost of consolidation is $O(L)$.

Returning to `deleteMin`, let us check its credit-potential invariant.

$$\begin{aligned} &\textsc{Cost} \leq 1 + t_0 + D(n), \qquad \textsc{Charge} = 2 + 2D(n),\\ &\textsc{Credit} \geq 1 + D(n) - t_0,\\ &\Delta\Phi \leq 1 + D(n) - t_0. \end{aligned}$$

We need to explain our bound for $\Delta\Phi$. Let t_0, m_0 refer to the values of $t(H)$ and $m(H)$ before this `deleteMin` operation. If Φ_0, Φ_1 are (respectively) the potentials before and after this operation, then $\Phi_0 = t_0 + 2m_0$ and $\Phi_1 \leq 1 + D(n) + 2m_0$. To see this bound on Φ_1, note that no node can have degree more than $D(n)$ (by definition of $D(n)$) and hence there are at most $1 + D(n)$ trees after consolidation. Moreover, there are at most m_0 marked after consolidation. Then $\Delta\Phi = \Phi_1 - \Phi_0 \leq 1 + D(n) - t_0$, as desired.

`decreaseKey`(x, k, H): this is the remaining operation and we will exploit the marking of items in a crucial way. First, we cut x iff x is not a root. Now x is in the root-list, so we can freely decrease the key of x to k. We need to update $H.\texttt{Min}$, etc. If x was marked, it is now unmarked. If x was not cut, this terminates the process. Otherwise, let y be the parent of x. If y was unmarked and y is not a root, we now mark y. But suppose y was marked (*i.e.*, has previously lost a child). Then we are suppose to cut y and recursively check if the parent of y was cut, and so on. We call this the "cascading cut of y" and captured it using the following fragment of code:

```
CascadingCut(y, H):
    if (y.mark = false and y ≠ root) then y.mark := true;
    if y ≠ root then
            Cut(y, H);
            CascadingCut(y.Parent, H).
```

Note that if $c \geq 1$ is the number of cuts, then $t(H)$ is increased by c, but $m(H)$ is decreased by $c-1$ or c (the latter iff x was marked). This implies $\Delta\Phi \leq c - 2(c-1) = 2 - c$. If

$$\textsc{Cost} \leq c, \quad \textsc{Charge} = 2, \quad \textsc{Credit} \geq 2 - c,$$

then the credit-potential invariant is verified.

SUMMARY: we have achieved our goal of charging $O(1)$ units to every operation except for `deleteMin` which is charged $O(1) + D(n)$. We next turn to bounding $D(n)$.

§7. Degree Bound

Our goal is to show that $D(n) = O(\log n)$.

Recall the ith Fibonacci number $i = 0, 1, 2, \ldots$ is defined by $F_i = i$ if $i = 0, 1$ and $F_i = F_{i-1} + F_{i-2}$ for $i \geq 2$. Thus the sequence of Fibonacci numbers starts out as

$$0, 1, 1, 2, 3, 5, 8, \ldots.$$

We will use two simple facts:

(a) $F_i = 1 + \sum_{j=1}^{i-2} F_j$ for $i \geq 2$.

(b) $F_{j+2} \geq \phi^j$ for $j \geq 0$, where $\phi = (1 + \sqrt{5})/2 > 1.618$.

Fact (a) follows easily by induction, or better still, by "unrolling" the recurrence for F_i. For fact (b), we observe that ϕ is a solution to the equation $x^2 - x - 1 = 0$ so $\phi^2 = 1 + \phi$. Clearly $F_2 = 1 \geq \phi^0$ and $F_3 = 2 \geq \phi^1$. Inductively,

$$F_{j+2} = F_{j+1} + F_j \geq \phi^{j-1} + \phi^{j-2} = \phi^{j-2}(\phi + 1) = \phi^j.$$

Let x be a node in a Fibonacci heap with n items, and let

$$y_1, y_2, \ldots, y_d \tag{6}$$

be the children of x, given in the order in which they are linked to x. So $x.\texttt{degree} = d$ and y_1 is the earliest child (among $y_1, \ldots, y_d$) to be linked to x.

Lemma 5

$$y_i.\texttt{degree} \geq \begin{cases} 0 & \text{if } i = 1 \\ i - 2 & \text{if } i \geq 2 \end{cases}$$

Proof. This is clearly true for $i = 1$. For $i \geq 2$, note that when y_i was linked to x, the degree of x is at least $i - 1$ (since at least $y_1, \ldots, y_{i-1}$ are children of x at the moment of linking). Hence, the degree of y_i at that moment is at least $i - 1$. But we allow y_i to lose at most one child before cutting y_i. Since y_i is not cut from x, the degree of y_i is at least $i - 2$. **Q.E.D.**

Lemma 6 *Let* Size(x) *denote the number of nodes in the subtree rooted at x and d the degree of x. Then*

$$\text{Size}(x) \geq F_{2+d}, \quad d \geq 0.$$

Proof. This is seen by induction on Size(x). The result is true when Size(x) $= 1, 2$ since in these cases $d = 0, 1$, respectively. If Size(x) ≥ 3, let $y_1, \ldots, y_d$ be the children of x as in (6). Then

$$\begin{aligned} \text{Size}(x) &= 1 + \sum_{i=1}^{d} \text{Size}(y_i) \\ &\geq 1 + \sum_{i=1}^{d} F_{y_i.\texttt{degree}+2} \text{ (by induction)} \\ &\geq 2 + \sum_{i=2}^{d} F_i \text{ (by last lemma)} \\ &= 1 + \sum_{i=1}^{d} F_i = F_{d+2}. \end{aligned}$$

Q.E.D.

It follows that if x has degree d, then

$$n \geq \text{Size}(x) \geq F_{d+2} \geq \phi^d.$$

Taking logarithms, we immediately obtain:

Lemma 7

$$D(n) \leq \log_\phi(n).$$

This completes our analysis of Fibonacci heaps. It is now clear why the name "Fibonacci" arises.

Exercises

Exercise 7.1: Suppose that instead of cutting a node just as it is about to lose a second child, we cut a node just as it is about to lose a third child. Carry out the analysis as before. Discuss the pros and cons of this variant Fibonacci heap. ♢

Exercise 7.2:

(a) Determine $\widehat{\phi}$, the other root of the equation $x^2 - x - 1 = 0$. Numerically compute $\widehat{\phi}$ to 3 decimal places.

(b) Determine F_i exactly in terms of ϕ and $\widehat{\phi}$ HINT: $F_i = A\phi^i + B\widehat{phi}^i$ for constants A, B.

(b) What is the influence of the $\widehat{\phi}$-term on the relative magnitude of F_i? ♢

End Exercises

§8. Pointer Model of Computation

There is an esthetically displeasing feature in our consolidation algorithm, namely, its use of array indexing does not seem to conform to the style used in the other operations. Intuitively, unlike the other operations, indexing does not fit within the "pointer model" of computation. It is instructive to formalize such a model.

A **pointer program** Π consists of a finite sequence of instructions that operate on an implicit potentially infinite digraph G. All program variables in Π are of type pointer, but we also manipulate integer values via these pointers. Each pointer points to some node in G. Each node N in G has four components:

(integer-value, 0-pointer, 1-pointer, 2-pointer).

These are accessed as P.`Val`, P.0, P.1 and P.2 where P is any pointer variable that points to N. There is a special node $N_0 \in G$ and this is pointed to by the nil pointer. By definition. nil.`Val` $= 0$ and nil.i = nil for $i = 0, 1, 2$. Note that with 3 pointers, it is easy to model binary trees.

Pointer expressions. In general, we can specify a node by a **pointer expression**, ⟨`pointer-expr`⟩, which is either the constant nil, the New() operator, or has the form $P.w$ where P is a pointer variable and $w \in \{0,1,2\}^*$. The string w is also called a **path**. Examples of pointer expressions:

$$\mathsf{nil}, \textsc{New}(), P, P.0, P.1, P.2, P.1201, P.212012l$$

where P is a pointer variable. The New() operator (with no arguments) returns a returns a pointer to a "spanking new node" N where $N.0 = N.1 = N.2 = \mathsf{nil}$ and $N.\texttt{Val} = 1$. The only way to access a node or its components is via such pointer expressions.

The integer values stored in nodes are unbounded and one can perform the four arithmetic operations; compare two integers; and assign to an integer variable from any integer expression (see below).

We can compare two pointers for equality or inequality, and can assign to a pointer variable from another pointer variable or the constant nil or the function New(). Assignment to a nil pointer has no effect. Note that we are not allowed to do pointer arithmetic or to compare them for the "less than" relation.

The assigment of pointers can be explained with an example:

$$P.0121 \leftarrow Q.20002$$

If N is the node referenced by $P.012$ and N' is the node referenced by $Q.20002$, then we are setting $N.1$ to point to N'. If N is the nil node, then this assignment has no effect.

Naturally, we use the result of a comparison to decide whether or not to branch to a labelled instruction. Assume some convention for input and output. For instance, we may have two special pointers P_{in} and P_{out} that point (respectively) to the input and output of the program.

To summarize:

a pointer program is a sequence of instructions (with an optional label) of the following type:

- Value Assignment: ⟨`pointer-expr`.Val⟩ ← ⟨`integer-expr`⟩;
- Pointer Assignment: ⟨`path-expr`⟩ ← ⟨`pointer-expr`⟩;
- Pointer Comparison: if ⟨`pointer-expr`⟩ = ⟨`pointer-expr`⟩ then goto ⟨`label`⟩;
- Value Comparison: if ⟨`integer-expr`⟩ ≥ 0 then goto ⟨`label`⟩;
- Halt

Integer expressions denote integer values. For instance

$$(74 * P.000) - (Q.21 + P)$$

where P, Q are pointer variables. Here, $P.000, Q.21, P$ denotes the values stored at the corresponding nodes. Thus, an integer expression ⟨`integer-expr`⟩ is either

- Base Case: any literal integer constant (e.g., $0, 1, 74, -199$), a ⟨`pointer-expr`⟩ (e.g., $P.012, Q$, nil); or
- Recursively:

$$(\langle\texttt{integer-expr}\rangle\langle op\rangle\langle\texttt{integer-expr}\rangle)$$

where $\langle op \rangle$ is one of the four arithmetic operations. Recall that nil.Val $= 0$. Some details about the semantics of the model may be left unspecified for now. For instance, if we divide by 0, the program may be assumed to halt instantly.

For complexity modeling, we may assume each of the above operations take unit time regardless of the pointers or the size of the integers involved. Likewise, the space usage can be simplified to just counting the number of nodes used.

One could embellish it with higher level constructs such as while-loops. Or, we could impoverish it by restricting the integer values to Boolean values (to obtain a better accounting of the bit-complexity of such programs). In general, we could have pointer models in which the value of a node P.Val comes from any domain. For instance, to model computation over a ring R, we let P.Val be an element of R. We might wish to have an inverse to New(), to delete a node.

List reversal example. Consider a pointer program to reverse a singly-linked list of numbers (we only use 0-pointer of each node to point to the next node). Our program uses the pointer variables P, Q, R and we write $P \leftarrow Q \leftarrow R$ to mean the sequential assignments "$P \leftarrow Q;\quad Q \leftarrow R;$".

```
ReverseList:
Input: P_in, pointer to a linked list.
Output: P_out, pointer to the reversal of P_in.
      P ← nil; Q ← P_in;
      if Q = nil then goto E;
           R ← Q.0 ← P;
L:    if R = nil then goto E;
T:         P ← Q ← R ← Q.0 ← P;
      goto L;
E:    P_out ← Q.
```

This program is easy to understand once the invariant assertion preceding line T is understood (see Figure 9 and exercise).

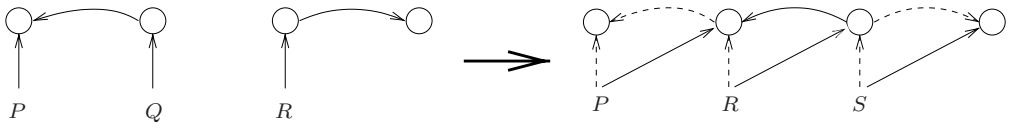

Figure 9: The transformation at line T.

Remark: This model may be more convenient than Turing machines to use as a common basis for discussing complexity theory issues. The main reservation comes from our unit cost for unbounded integers operations. In that case we can either require that all integers be bounded, or else charge a suitable cost $M(n)$ for multiplying n-bit integers, etc, reflecting the Turing machine cost. Of course, the use of pointers is still non-elementary from the viewpoint of Turing machines, but this is precisely the convenience we gain.

Pointer based consolidation. We outline a purely pointer-based method for consolidation: assume that if $k \leq D(n)$ is the maximum degree of any node (past or present) in the Fibonacci heap, we have a doubly-

linked list of nodes

$$(R_0, R_1, \ldots, R_k).$$

We call this the "degree register" because every node in the heap of degree i will have a pointer to R_i. Here k is the largest degree of a node that has been seen so far. Note that when we link x to y then the degree of y increments by one and when we cut x, then the parent of x decrements by one, and these are the only possibilities. If item x has its degree changed from i to $i \pm 1$ then we can re-register x by pointing it to $R_{i\pm1}$ in constant time. Occasionally, we have to extend the length of the register by appending a new node R_{k+1} to the doubly-linked list (when some node attains a degree $k+1$ that is larger than any seen so far). It is thus easy to maintain this degree register. Now suppose we must consolidate a root list J. By going through the items in J, we can create (with the help of the degree register) a list of lists

$$(L_0, L_1, \ldots, L_k)$$

where list L_i comprises the roots of degree i in J. This takes $O(D(n) + t)$ operations if J has t elements. It is now easy to consolidate the lists $L_0, \ldots, L_k$ into one list in which no two trees have the same degree, using $O(t)$ time. The cost of this procedure is $O(D(n) + t)$, as in the solution that uses array indexing.

Exercises

Exercise 8.1: State the invariant before line T in the pointer reversal program; then proving the program correct. ♢

Exercise 8.2: Write the pointer program for the consolidation. ♢

Exercise 8.3: Implement in detail all the Fibonacci heap algorithms using our pointer model, ♢

Exercise 8.4: Write a sorting program and a matrix multiplication program in this model. What is the time complexity of your algorithms? ♢

End Exercises

§9. Application to Minimum Spanning Tree

An original application of Fibonacci heaps is to compute minimum spanning trees (MST). This was introduced in Lecture IV §4. We now consider Prim's algorithm for MST, using the notations of §IV.4. So, the input is a connected bigraph $G = (V, E; C)$ with cost function $C : E \to \mathbb{R}$.

Although our goal is to compute a minimum spanning tree, let us simplify our task by computing only the **cost** of a minimum spanning tree. This is consistent with a general point of pedagogy: for many computational problems that seek to compute a data structure $D = D^*$ which minimizes an associated cost function $f(D)$, it is easier to just maintain $f(D)$ than to maintain D. Furthermore, we could subsequently indicate additional book-keeping steps to transform the algorithm that produces the minimum cost $f(D^*)$ into an algorithm that produces the optimal data-structure D^*.

Prim-safe sets. It is easy to see that if U is a singleton then U is Prim-safe. Suppose U is Prim-safe and we ask how U might be extended to a larger Prim-safe set. Let us maintain the following information about U:
i) $\texttt{mst}[U]$, denoting the cost of the minimum spanning tree of $G|U$.
ii) For each $v \in V - U$, the least cost $\texttt{lc}_U[v]$ of an edge connecting v to U:

$$\texttt{lc}_U[v] := \min\{C(v,u) : (v,u) \in E, u \in U\}.$$

We usually omit the subscript U and just write "$\texttt{lc}[v]$" without confusion.

In order to find a node $u^* \in V - U$ with the minimum $\texttt{lc}$-value, we will maintain $V - U$ as a **single**[6] mergeable queue Q in which the least cost $\texttt{lc}[u]$ serves as the key of the node $u \in V - U$. Hence extending the Prim-safe set U by a node u^* amounts to a `deleteMin` from the mergeable queue. After the deletion, we must update the information $\texttt{mst}[U]$ and $\texttt{lc}[v]$ for each $v \in V - U$. But we do not really need to consider every $v \in V - U$: we only need to update $\texttt{lc}[v]$ for those v that are adjacent to u^*. The following code fragment captures our intent.

```
UPDATE(u*, U):
1.  U ← U ∪ {u*}.        {This step need not be performed}
2.  mst[U] ← mst[U] + lc[u*].
3.  for v adjacent to u* and v ∉ U, do
        if lc[v] > C[v, u*] then
            lc[v] ← C[v, u*]}.
            DecreaseKey(v, lc[v], Q).
```

We need not explicitly carry out step 1 because U is implicitly maintained as the complement of the items in Q. We now present the MST Cost version of Prim's algorithm.

```
MST COST ALGORITHM:
    Input: G = (V, E; C), a connected costed bigraph.
    Output: the cost of an MST of G.
    INITIALIZE:
    1.  U ← {v0}; mst[U] ← 0;
    2.  for v ∈ V − U, do lc[v] ← C(v, v0);
    3.  Set up V − U as a single mergeable queue Q:
            Q ← MakeQueue();
            Insert each element of V − U into Q.
    LOOP:
    4.  while Q ≠ ∅, do
            u* ← deleteMin(Q);
            UPDATE(u*, U).
    5.  return(mst[U]).
```

We do not need to maintain U explicitly, although it seems clearer to put this into our pseudo-code above. In practice, the updating of U can be replaced by a step to add edges to the current MST.

[6] So we are not using the full power of the mergeable queue ADT which can maintain several mergeable queues. In particular, we never perform the union operation in this application.

Analysis. The correctness of this algorithm is immediate from the preceding discussion. To bound its complexity, let $n := |V|$ and $m := |E|$. Assume that the mergeable queue is implemented by a Fibonacci heap. In the Update subroutine, updating the value of `lc`$[v]$ becomes a DecreaseKey operation. Each operation in Update can be charged to an edge or a vertex. As each edge or vertex is charged at most once, and since the amortized cost of each operation is $O(1)$, the cost of all the updates is $O(m+n)$. The initialization takes $O(n)$ time. In the main procedure, we make $n-1$ passes through the whileloop. So we perform $n-1$ `deleteMin` operations, and as the amortized cost is $O(\log n)$ per operation, this has total cost $O(n \log n)$. We have proven:

Theorem 8 *The cost of a minimum spanning tree of a graph $(V, E; C)$ can be found in $O(|V| \log |V| + |E|)$ operations.*

Final Remarks. The amortization idea is closely related to two other topics. One is "self-organizing data structures". Originally, this kind of analysis is undertaken by assuming the input has certain probability distribution. McCabe (1965) is the first to discuss the idea of move-to-front rule. See "An account of self-organizing systems", W.J. Hendricks, *SIAM J.Comp.*, 5:4(1976); also "Heuristics that dynamically organizes data structures", James R. Bitner, *SIAM J.Comp.*, 8:1(1979)82-100. But starting from the work of Sleator and Tarjan, the competitive analysis approach has become dominant. Albers and Westbrook gives a survey in [2]. Indeed, competitive analysis is the connection to the other major topic, "online algorithms". Albers gives a survey [1].

Exercises

Students should be able to demonstrate understanding of Prim's algorithm by doing hand simulations. The first exercise illustrates a simple tabular form for hand simulation.

Exercise 9.1: Hand simulate Prim's algorithm on the following graph (figure 10) beginning with v_1:

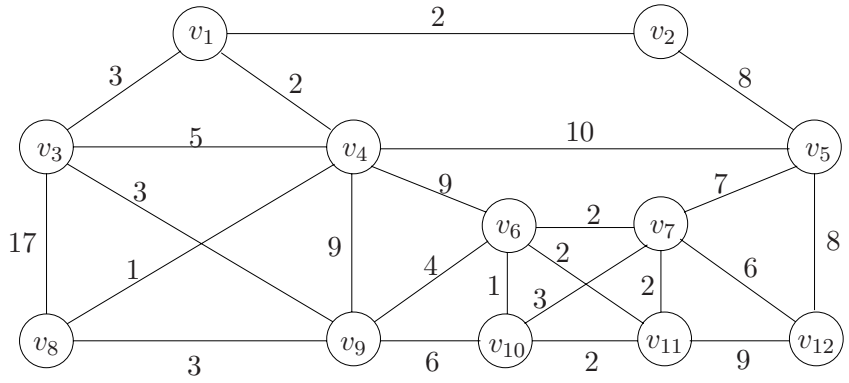

Figure 10: Graph of a House

It amounts to filling in the following table, row by row. We have filled in the first two rows already.

i	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	$mst[U]$	New Edge
1	$\underline{2}$	3	2	∞	∞	∞	∞	∞	∞	∞	∞	2	(v_1, v_2)
2	*	"	$\underline{"}$	8	"	"	"	"	"	"	"	4	(v_1, v_4)

Note that the

minimum cost in each row is underscored, indicating the item to be removed from the priority queue. ♢

Exercise 9.2: Let G_n be the graph with vertices $\{1, 2, \ldots, n\}$ and for $1 \leq i < j \leq n$, we have an edge (i, j) iff i divides j. For instance, $(1, j)$ is an edge for all $1 < j \leq n$. The **cost** of the edge (i, j) is $j - i$.
(a) Hand simulate (as in the previous exercise) Prim's algorithm on G_{10}. Show the final MST and its cost.

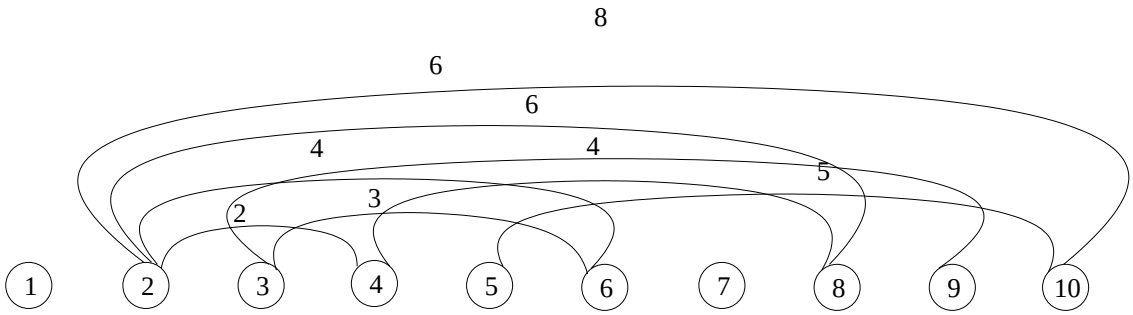

Figure 11: G_{10}: edges from node 1 are omitted for clarity.

(b) What can you say about the MST of G_n? Is it unique? What is the asymptotic cost of the MST? ♢

Exercise 9.3: Modify the above algorithm to compute a minimum spanning tree. ♢

Exercise 9.4: Modify the above algorithm to compute a minimum spanning forest in case the input graph is not connected. ♢

Exercise 9.5: Let $G = (V, E; \mu)$ be an edge-costed bigraph and $S \subseteq E$, $U \subseteq V$. Let $V(S) = \{v \in V : \exists u, (u, v) \in S\}$ denote the *vertices of* S, and $G|U := (U, E'; \mu)$ where $E' = E \cap \binom{U}{2}$ denote the *restriction of* G *to* U. We define S to be *prim-safe* if S is an MST of $G|V(S)$ and S can be extended into an MST of G. We define U to be *prim-safe* if U is singleton or there exists a prim-safe set S of edges such that $U = V(S)$. Show or give a counter-example:
(a) S is a tree of $G|V(S)$ and can be extended into an MST of G implies S is prim safe.
(b) U is prim-safe implies every MST of $G|U$ is prim-safe. ♢

End Exercises

§A. APPENDIX: List Update Problem

The splay tree idea originates in the "move-to-front rule" heuristic for following **list update problem**: let L be a doubly-linked list of **items** where each item has a unique key. For simplicity, we usually write L as a sequence of keys. This list supports the **access request**. Each access request r is specified by a key (also denoted r), and we satisfy this request by returning a pointer to the item in L with key r. (We assume such an item always exist.) We are interested in a special class of algorithms: such an algorithm α, on an input L and r, searches sequentially in L for the key r by starting at the head of the list. Upon finding the item with key r, α is allowed to move the item to some position nearer the head of the list (the relative ordering of the other items is unchanged). Here are three alternative rules which specify the new position of an updated item:

- (R_0) The **lazy rule** never modifies the list L.
- (R_1) The **move-to-front** rule always make updated item the new head of the list L.
- (R_2) The **transpose rule** just moves the updated item one position closer to the head of the list.

Let α_i denote the list update algorithm based on Rule R_i ($i = 0, 1, 2$). For instance, α_1 is the "move-to-front algorithm". For any algorithm α, let $COST_\alpha(r, L)$ denote the cost of an update request r on a list L using α. For $i = 0, 1, 2$, we write $COST_i(r, L)$ instead of $COST_{\alpha_i}(r, L)$. We may define $COST_i(r, L)$ to be $1 + j$ where j is the position of the accessed item in L. If α is an update algorithm, then $\alpha(L, r)$ denotes the updated list upon applying α to L, r. We extend this notation to a sequence $U = \langle r_1, r_2, \ldots, r_n\rangle$ of requests, by defining

$$\alpha(L, U) := \alpha(\alpha(L, \langle r_1, \ldots, r_{n-1}\rangle), r_n).$$

Similarly, $COST_\alpha(L, U)$ or $COST_i(L, U)$ denotes the sum of the individual update costs.

Example: Let $L = \langle a, b, c, d, e\rangle$ be a list and c an update request. Then $\alpha_0(L, c) = L$, $\alpha_1(L, c) = \langle c, a, b, d, e\rangle$ and $\alpha_2(L, c) = \langle a, c, b, d, e\rangle$. Also $COST_i(L, c) = 4$ for all $i = 0, 1, 2$.

Probabilistic Model. We analyze the cost of a sequence of updates under the lazy rule and the move-to-front rule. We first analyze a probabilistic model where the probability of updating a key k_i is p_i, for $i = 1, \ldots, m$. The lazy rule is easy to analyze: if the list is $L = \langle k_1, \ldots, k_m\rangle$ then the expected cost of a single access request is

$$C(p_1, \ldots, p_m) = \sum_{i=1}^{m} i \cdot p_i.$$

It is easy to see that this cost is minimized if the list L is rearranged so that $p_1 \geq p_2 \geq \cdots \geq p_m$; let C^* denote this minimized value of $C(p_1, \ldots, p_m)$.

What about the move-to-front rule? Let $p(i, j)$ be the probability that k_i is in front of k_j in list L. This is the probability that, if we look at the last time an update involved k_i or k_j, the operation involves k_i. Clearly

$$p(i, j) = \frac{p_i}{p_i + p_j}.$$

The expected cost to update k_i is

$$1 + \sum_{j=1, j\neq i}^{m} p(j, i).$$

The expected cost of an arbitrary update is

$$\begin{aligned}
\widehat{C} &:= \sum_{i=1}^{m} p_i \left[1 + \sum_{j=1, j\neq i}^{m} p(i,j) \right] \\
&= 1 + \sum_{i=1}^{m} \sum_{j\neq i}^{m} p_i \cdot p(i,j) \\
&= 1 + 2 \sum_{1\leq j<i\leq m} \frac{p_i p_j}{p_i + p_j} \\
&= 1 + 2 \sum_{i=1}^{m} p_i \sum_{j=1}^{i-1} \cdot p(j,i) \\
&\leq 1 + 2 \sum_{i=1}^{m} p_i \cdot (i-1) \\
&= 2C^* - 1.
\end{aligned}$$

This proves

$$\widehat{C} < 2C^*. \tag{7}$$

Amortization Model. Let us now consider the amortized cost of a sequence of updates

$$U = (r_1, r_2, \ldots, r_n) \tag{8}$$

on an initial list L_0 with m items. Clearly the worst case cost per update is $O(m)$. So, updates over the sequence U costs $O(mn)$. This worst case bound cannot be improved if we use the lazy rule. The best case for the lazy rule is $O(1)$ per update, or $O(n)$ overall.

What about the move-to-front rule? In analogy to equation (7), we show that it is never incur more than twice the cost of any update algorithm. In particular, it is never more than twice cost of an optimal offline update algorithm α_*: if the cost of α_* is denoted $COST_*$, we prove

$$COST_1(L, U) \leq 2 \cdot COST_*(L, U). \tag{9}$$

We use an amortization argument based on potential functions. A pair (k, k') of keys is an **inversion** in a pair (L, L') of lists if k occurs before k' in L but k occurs after k' in L'. Fix a list L_0 and for any list L, define its **potential** $\Phi(L)$ to be the number of inversions in (L_0, L).

Consider the jth request ($j = 1, \ldots, n$). Let L_j (resp. L_j^*) be the list produced by the move-to-front (resp. optimal) algorithm after the jth request. Write Φ_j for $\Phi(L_j)$. Let c_j and c_j^* denote the cost of serving the jth request under two algorithms (respectively). Let x_j be the item accessed in the jth request and k_j is the number of items that are in front of x_j in both lists L_j and L_j^*. Let ℓ_j be the number of items that are in front of x_j in L_j but behind x_j in L_j^*. Hence

$$c_j = k_j + \ell_j + 1, \quad c_j^* \geq k_j + 1.$$

On the other hand, the number of inversions destroyed is ℓ_j and the number of inversions created is at most k_j. It follows

$$\Phi_j - \Phi_{j-1} \leq k_j - \ell_j.$$

Combining these two remarks,

$$\begin{aligned}
c_j + \Phi_j - \Phi_{j-1} &\leq 2k_j + 1 \\
&\leq 2c_j^* - 1.
\end{aligned}$$

Summing up over all $j = 1, \ldots, n$, we obtain

$$\begin{aligned} COST_1(L_0, U) &= \left(\sum_{j=1}^{n} c_j\right) + \Phi_n - \Phi_0 \\ &\leq \sum_{j=1}^{n} (2c_j^* - 1), \qquad (\text{since } \Phi_n \geq 0, \Phi_0 = 0) \\ &= 2COST_*(L_0, U) - n. \end{aligned}$$

Competitive Algorithms. Let $\beta(k)$ be a function of k. We say an algorithm α is $\beta(k)$-**competitive** if there is some constant a, for all input lists L of length k and for all sequences U of requests

$$COST_\alpha(L, U) \leq \beta(k) \cdot COST_*(L, U).$$

Here $COST_*$ is the cost incurred by the optimal offline algorithm.

We have just shown that the Move-to-Front algorithm is 2-competitive. This idea of competiveness from Sleator and Tarjan is an extremely powerful one as it opens up the possibility of measuring the performance of online algorithms (such as the move-to-front algorithm) without any probabilistic assumption on the input requests.

Remark. An application of the list update problem is data-compression (Exercise). Chung, Hajela and Seymour [3] determine that cost of the move-to-front rule over the cost of an optimal static ordering of the list (relative to some probability of accessing each item) is $\pi/2$. See also Lewis and Denenberg [4] and Purdom and Brown [6].

Exercises

Exercise A.1: We extend the list update problem above in several ways:
(a) One way is to allow other kinds of requests. Suppose we allow insertions and deletions of items. Assume the following algorithm for insertion: we put the new item at the end of the list and perform an access to it. Here is the deletion algorithm: we access the item and then delete it. Show that the above analyses extend to a sequence of access, insert and delete requests.
(b) Extend the list update analysis to the case where the requested key k may not appear in the list.
(c) A different kind of extension is to increase the class of algorithms we analyze: after accessing an item, we allow the algorithm to to transpose any number of pairs of adjacent items, where each transposition has unit cost. Again, extend our analyses above. ♢

Exercise A.2: The above update rules R_i $(i = 0, 1, 2)$ are memoryless. The following two rules require memory.

- (R_3) The **frequency rule** maintains the list so that the more frequently accessed items occur before the less frequently accessed items. This algorithm, of course, requires that we keep a counter with each item.
- (R_4) The **timestamp rule** (Albers, 1995) says that we move the requested item x in front of the first item y in the list that precedes x and that has been requested at most once since the last request to x. If there is no such y or if x has not been requested so far, do not move x.

(a) Show that R_3 is not c-competitive for any constant c.
(b) Show that R_4 is 2-competitive. ♢

Exercise A.3: (Bentley, Sleator, Tarjan, Wei) Consider the following data compression scheme based on any list updating algorithm. We encode an input sequence S of symbols by each symbol's position in a list L. The trick is that L is dynamic: we update L by accessing each of the symbols to be encoded. We now have a string of integers. To finally obtain a binary string as our output, we encode this string of integers by using a prefix code for each integer. In the following, assume that we use the move-to-front rule for list update. Furthermore, we use the prefix code of Elias in Exercise IV.1.1.6 that requires only

$$f(n) = 1 + \lfloor \lg n \rfloor + 2 \lfloor \lg(1 + \lg n) \rfloor$$

bits to encode an integer n.
(a) Assume the symbols are a, b, c, d, e and the initial list is $L = (a, b, c, d, e)$. Give the integer sequence corresponding to the string $S = abaabcdabaabecbaadae$. Also give the final binary string corresponding to this integer sequence.
(b) Show that if symbol x_i occurs $m_i \geq 0$ times in S then these m_i occurrences can be encoded using a total of

$$m_i f(m/m_i)$$

bits where $|S| = m$. HINT: If the positions of x_i in S are $1 \leq p_1 < p_1 < \cdots < p_{m_i} \leq m$ then the jth occurrence of x_i needs at most $f(p_j - p_{j-1})$. Then use Jensen's inequality for the concave function $f(n)$.
(c) If there are n distinct symbols $x_1, \ldots, x_n$ in S, define

$$A(S) := \sum_{i=1}^{n} \frac{m_i}{m} f\left(\frac{m}{m_i}\right).$$

Thus $A(S)$ bounds the average number of bits per symbol used by our compression scheme. Show that

$$A(S) \leq 1 + H(S) + 2 \lg(1 + H(S))$$

where

$$H(S) := \sum_{i=1}^{n} \frac{m_i}{m} \lg\left(\frac{m}{m_i}\right).$$

NOTE: $H(S)$ is the "empirical entropy" of S. It corresponds to the average number of bits per symbol achieved by the Huffman code for S. In other words, this online compression scheme achieves close to the compression of the offline Huffman coding algorithm. ♢

End Exercises

References

[1] S. Albers. Competitive online algorithms. BRICS Lecture Series LS-96-2, BRICS, Department of Computer Science, University of Aarhus, September 1996.

[2] S. Albers and J. Westbrook. A survey of self-organizing data structures. Research Report MPI-I-96-1-026, Max-Planck-Institut für Informatik, Im Stadtwald, D-66123 Saarbrücken, Germany, October 1996.

[3] F. R. K. Chung, D. J. Hajela, and P. D. Seymour. Self-organizing sequential search and hilbert's inequalities. *ACM Symp. on Theory of Computing*, 7, 1985. Providence, Rhode Island.

[4] H. R. Lewis and L. Denenberg. *Data Structures and their Algorithms.* Harper Collins Publishers, New York, 1991.

[5] A. Moffat, G. Eddy, and O. Petersson. Splaysort: Fast, versatile, practical. *Software - Practice and Experience*, 126(7):781–797, 1996.

[6] J. Paul Walton Purdom and C. A. Brown. *The Analysis of Algorithms.* Holt, Rinehart and Winston, New York, 1985.

[7] M. Sherk. Self-adjusting k-ary search trees. In *Lecture Notes in Computer Science*, volume 382, pages 373–380, 1989. Proc. **Workshop on Algorithms and Data Structures**, Aug. 17-19, 1989, Carleton University, Ottawa, Canada.

[8] D. D. Sleator and R. E. Tarjan. Self-adjusting binary search trees. *J. of the ACM*, 32:652–686, 1985.

[9] R. E. Tarjan. Amortized computational complexity. *SIAM J. on Algebraic and Discrete Methods*, 6:306–318, 1985.